



Volume 105
Number 2

May 2013

Published quarterly
by the
American Psychological
Association

ISSN 0022-0663

Journal of Educational Psychology

Arthur C. Graesser, *Editor*

Jill Fitzgerald, *Associate Editor*

David Francis, *Associate Editor*

Susan Goldman, *Associate Editor*

Young-Suk Kim, *Associate Editor*

Robert Klassen, *Associate Editor*

David N. Rapp, *Associate Editor*

Susan Sonnenschein, *Associate Editor*

Birgit Spinath, *Associate Editor*

Roman Taraban, *Associate Editor*

Jennifer Wiley, *Associate Editor*

Christopher Wolters, *Associate Editor*

www.apa.org/pubs/journals/edu

Marygrove College Library
8425 West McNichols Road
Detroit, MI 48221

Editor

Arthur C. Graesser, *University of Meuphis*

Associate Editors

Jill Fitzgerald, *University of North Carolina at Chapel Hill, Emeritus*
David Francis, *University of Houston*
Susan Goldman, *University of Illinois at Chicago*
Young-Suk Kim, *Florida State University*
Robert Klassen, *University of Alberta*
David N. Rapp, *Northwestern University*
Susan Sonnenschein, *University of Maryland*
Birgit Spinath, *University of Heidelberg, Heidelberg, Germany*
Roman Taraban, *Texas Tech University*
Jennifer Wiley, *University of Illinois at Chicago*
Christopher Wolters, *University of Houston*

Chief Editorial Assistant

Jean Edgar, *University of Memphis*

Advisory Editors

Mary D. Ainley, *University of Melbourne, Australia*
Shaaron Ainsworth, *University of Nottingham, United Kingdom*
Vincent Aleven, *Carnegie Mellon University*
Patricia Alexander, *University of Maryland, College Park*
Richard L. Allington, *University of Tennessee*
Ellen R. Altermatt, *Hanover College*
Ivan Ash, *Old Dominion University*
Carole Beal, *University of Arizona*
Hefer Bembenuddy, *Queens College, CUNY*
David A. Bergin, *University of Missouri, Columbia*
Daniel Bolt, *University of Wisconsin, Madison*
Mimi Bong, *Korea University, Seoul, Korea*
Julie L. Booth, *Temple University*
Lee Brantum-Martin, *Georgia State University*
M. Anne Britt, *Northern Illinois University*
Scott Brown, *University of Connecticut*
Eric S. Buhs, *University of Nebraska, Lincoln*
Adriana G. Bus, *Leiden University, The Netherlands*
Kirsten R. Butcher, *University of Utah*
Robert Calfee, *University of California, Riverside*
Martha Carr, *University of Georgia*
Kwangsung Cho, *University of Missouri, Columbia*
Timothy Cleary, *University of Wisconsin, Milwaukee*
Anne E. Cook, *University of Utah*
Kai Cortina, *University of Michigan*
Jennifer Cromley, *Temple University*
H. Michael Crowson, *University of Oklahoma*
Anne E. Cunningham, *University of California, Berkeley*
Teresa K. DeBacker, *The University of Oklahoma*
Sidney D'Mello, *University of Notre Dame*
John Dunlosky, *Kent State University*
Amanda M. Durik, *Northern Illinois University*
Gary Feng, *Educational Testing Service*
J. D. Fletcher, *Institute for Defense Analyses*
Lynn S. Fuchs, *Vanderbilt University*
Linda Gambrell, *Clemson University*
James P. Gee, *Arizona State University*
Arthur M. Glenberg, *Arizona State University*
Adele E. Gottfried, *California State University*
Steve Graham, *Arizona State University*
Barbara A. Greene, *University of Oklahoma*
John Guthrie, *University of Maryland*
Douglas Hacker, *University of Utah*
Vernon C. Hall, *Syracuse University*
Jill Hamm, *University of North Carolina, Chapel Hill*
John Hattie, *University of Auckland, New Zealand*
Mary Hegarty, *University of California, Santa Barbara*
Jan N. Hughes, *Texas A&M University*
Slava Kalyuga, *University of South Wales, Australia*
Avi Kaplan, *Temple University*
Michael J. Kieffer, *New York University*
Beth Kurtz-Costes, *University of North Carolina, Chapel Hill*
Dan Lapsley, *University of Notre Dame*
Willy Lens, *University of Leuven, Belgium*
Elizabeth A. Linnenbrink-Garcia, *Duke University*
Robert Lorch, *University of Kentucky*
Joseph P. Magliano, *Northern Illinois University*
Andrew Martin, *University of Sydney, Australia*
Andrew J. Mashburn, *Portland State University*
Linda Mason, *Pennsylvania State University*
Richard E. Mayer, *University of California, Santa Barbara*
Charles MacArthur, *University of Delaware*
Catherine McBride-Chang, *The Chinese University of Hong Kong, China*
Nicole M. McNeil, *University of Notre Dame*
Debra K. Meyer, *Elmhurst College*
Keith Millis, *Northern Illinois University*
Alexandre J. S. Morin, *University of Western Sydney, Australia*
Tamera B. Murdock, *University of Missouri, Kansas City*
P. Karen Murphy, *Pennsylvania State University*
Benjamin Nagengast, *Eberhard Karls University of Tübingen*
Mitchell J. Nathan, *University of Wisconsin, Madison*
E. Michael Nussbaum, *University of Nevada, Las Vegas*
Rollanda E. O'Connor, *University of California, Riverside*
Harry O'Neil, *University of Southern California*
Tenaha O'Reilly, *Educational Testing Service*
Philip Parker, *University of Western Sydney, Australia*
Helen Patrick, *Purdue University*
Erika Patall, *University of Texas, Austin*
Reinhard Pekrun, *University of Munich, Germany*
Yaacov Petscher, *Florida State University*
Gary Phye, *Iowa State University*
Keenan Pituch, *University of Texas, Austin*
Jan L. Plass, *New York University*

Patrick Proctor, *Boston College*
Katherine Rawson, *Kent State University*
Robert Renaud, *University of Manitoba, Canada*
Alexander Renkl, *University of Freiburg, Germany*
Catherine Richards-Tutor, *California State University, Long Beach*
Bethany Rittle-Johnson, *Vanderbilt University*
Daniel Robinson, *University of Texas, Austin*
Philip Rodkin, *University of Illinois at Urbana-Champaign*
Christopher A. Sanchez, *Arizona State University*
Katherine Scheiter, *Knowledge Media Research Center, Germany*
Marlene Schommer-Aikins, *Wichita State University*
Gregory Schraw, *University of Nevada, Las Vegas*
Dale Schunk, *University of North Carolina, Greensboro*
Christian D. Schunn, *University of Pittsburgh*
Paula J. Schwanenflugel, *University of Georgia*
Colleen M. Seifert, *University of Michigan*
Timothy Shanahan, *University of Illinois, Chicago*
Gale M. Sinatra, *University of Southern California*
Einar M. Skaalvik, *Norwegian University of Science and Technology, Norway*
John Sweller, *University of New South Wales, Australia*
Keith Thiede, *Boise State University*
Theresa A. Thorkildsen, *University of Illinois, Chicago*
Wendy Troop-Gordon, *North Dakota State University*
Chia-Wen Tsai, *Ming Chuan University-Taiwan*
Timothy Urdan, *Santa Clara University*
Ellen Usher, *University of Kentucky*
Regina Vollmeyer, *University of Frankfurt, Germany*
Jeffrey Walczyk, *Louisiana Technical University*
Charles A. Weaver III, *Baylor University*
Joanna P. Williams, *Columbia University*
Phil Winne, *Simon Fraser University, Canada*
Moshe M. Zeidner, *University of Haifa, Israel*

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Change of Address: Send change of address notice and a recent mailing label to the attention of Subscriptions Department, APA, or e-mail www.apa.org/pubs/journals/subscriptions.aspx 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee periodicals forwarding postage.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Microform Editions: For information regarding microform editions, write to University Microfilms, Ann Arbor, MI 48106.

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Art Graesser, Journal of Educational Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Copyright and Permission: Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA written permission to reproduce a journal article in full or journal text of more than 800 cumulative words or more than 3 tables and/or figures. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law; (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/13/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. For more information along with a permission form, go to www.apa.org/about/contact/copyright/index.aspx

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See <http://my.apa.org/access.html>.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Susan J. A. Harris, *Senior Director, Journals Program*; John Breithaupt, *Director, Journal Services*; Paige W. Jackson, *Director, Editorial Services*; Megan Mabe-Stanberry, *Account Manager*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

The **Journal of Educational Psychology**® (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2013 rates follow: *Nonmember Individual:* \$198 Domestic, \$227 Foreign, \$240 Air Mail. *Institutional:* \$687 Domestic, \$736 Foreign, \$751 Air Mail. *APA Member:* \$86. *APA Student Affiliate:* \$59. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of Educational Psychology® is a registered trademark of the American Psychological Association

Articles

© 2013
American
Psychological
Association

- 249 A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games
Pieter Wouters, Christof van Nimwegen, Herre van Oostendorp, and Erik D. van der Spek
- 266 Reducing Verbal Redundancy in Multimedia Learning: An Undesired Desirable Difficulty?
Carole L. Yue, Elizabeth Ligon Bjork, and Robert A. Bjork
- 278 Learning With Animation and Illusions of Understanding
Eugene S. Paik and Gregory Schraw
- 290 Explanation Feedback Is Better Than Correct Answer Feedback for Promoting Transfer of Learning
Andrew C. Butler, Namrata Godbole, and Elizabeth J. Marsh
- 299 Note-Taking With Computers: Exploring Alternative Strategies for Improved Recall
Dung C. Bui, Joel Myerson, and Sandra Hale
- 310 A (Pan-Canadian) Cluster Randomized Control Effectiveness Trial of the ABRACADABRA Web-Based Literacy Program
Robert Savage, Philip C. Abrami, Noella Piquette, Eileen Wood, Gia Deleveaux, Sukhbinder Sanghera-Sidhu, and Giovani Burgos
- 329 Do Films Make You Learn? Inference Processes in Expository Film Comprehension
Maike Tibus, Anke Heier, and Stephan Schwan
- 341 Managing Face Threats and Instructions in Online Tutoring
Benjamin Brummernhenrich and Regina Jucks
- 351 Extraneous Perceptual Information Interferes With Children's Acquisition of Mathematical Knowledge
Jennifer A. Kaminski and Vladimir M. Sloutsky
- 364 Complex Problem Solving in Educational Contexts—Something Beyond g: Concept, Assessment, Measurement Invariance, and Construct Validity
Samuel Greiff, Sascha Wüstenberg, Gyöngyvér Molnár, Andreas Fischer, Joachim Funke, and Benő Csapó
- 380 A Meta-Analysis of the Efficacy of Teaching Mathematics With Concrete Manipulatives
Kira J. Carbonneau, Scott C. Marley, and James P. Selig
- 401 Modeling Writing Development: Contribution of Transcription and Self-Regulation to Portuguese Students' Text Generation Quality
Teresa Limpo and Rui A. Alves
- 414 Do Early Literacy Skills in Children's First Language Promote Development of Skills in Their Second Language? An Experimental Evaluation of Transfer
J. Marc Goodrich, Christopher J. Lonigan, and JoAnn M. Farver
- 427 Enhancing a Brief Writing Intervention to Combat Stereotype Threat Among Middle-School Students
Natasha K. Bowen, Kate M. Wegmann, and Kristina C. Webber
- 436 A Contextualized View on Long-Term Predictors of Academic Performance
Janine Gut, Giselle Reimann, and Alexander Grob

- 444 The Effects of Single-Sex Compared With Coeducational Schooling on Mathematics and Science Achievement: Data From Korea
Erin Palilke, Janet Shibley Hyde, and Janet E. Mertz
- 453 The Transition From Informal to Formal Mathematical Knowledge: Mediation by Numeral Knowledge
David J. Purpura, Arthur J. Baroody, and Christopher J. Lonigan
- 465 Early Teacher Expectations Disproportionately Affect Poor Children's High School Performance
Nicole S. Sorhagen
- 478 Sex-Specific Differential Prediction of College Admission Tests: A Meta-Analysis
Franziska T. Fischer, Johannes Schult, and Benedikt Hell
- 489 The Internal/External Frame of Reference of Academic Self-Concept: Extension to a Foreign Language and the Role of Language of Instruction
Man K. Xu, Herbert W. Marsh, Kit-Tai Hau, Irene T. Ho, Alexandre J. S. Morin, and Adel S. Abduljabbar
- 504 The Role of Goal Attainment Expectancies in Achievement Goal Pursuit
Corwin Senko and Chris S. Hulleman
- 522 Constructing Motivation Through Choice, Interest, and Interestingness
Erika A. Patall
- 535 Effectiveness of the KiVa Antibullying Program: Grades 1–3 and 7–9
Antti Kärnä, Marinus Voeten, Todd D. Little, Erkki Alanen, Elisa Poskiparta, and Christina Salmivalli
- 552 Early Adolescent Depression Symptoms and School Dropout: Mediating Processes Involving Self-Reported Academic Competence and Achievement
Cintia V. Quiroga, Michel Janosz, Sherri Bisset, and Alexandre J. S. Morin

Other

- 309 Call for Papers: Special Issue Ethical, Regulatory, and Practical Issues in Telepractice
- 551 Correction to Kärnä et al. (2012)
- iii Instructions to Authors
- ii Subscription Order Form

ORDER FORM

Start my 2013 subscription to the ***Journal of Educational Psychology***® ISSN: 0022-0663

_____ \$86.00	APA MEMBER/AFFILIATE	_____
_____ \$198.00	INDIVIDUAL NONMEMBER	_____
_____ \$687.00	INSTITUTION	_____
	<i>In DC and MD add 6% sales tax</i>	_____
	TOTAL AMOUNT DUE	\$ _____

Subscription orders must be prepaid. Subscriptions are on a calendar year basis only. Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.



SEND THIS ORDER FORM TO

American Psychological Association
Subscriptions
750 First Street, NE
Washington, DC 20002-4242

Call **800-374-2721** or 202-336-5600
Fax **202-336-5568** :TDD/TTY **202-336-6123**
For subscription information,
e-mail: **subscriptions@apa.org**

☐ **Check enclosed** (make payable to APA)

Charge my: ☐ Visa ☐ MasterCard ☐ American Express

Cardholder Name _____

Card No. _____ Exp. Date _____

Signature (Required for Charge)

Billing Address

Street _____

City _____ State _____ Zip _____

Daytime Phone _____

E-mail _____

Mail To

Name _____

Address _____

City _____ State _____ Zip _____

APA Member # _____

EDUA13

A Meta-Analysis of the Cognitive and Motivational Effects of Serious Games

Pieter Wouters, Christof van Nimwegen,
and Herre van Oostendorp
Utrecht University

Erik D. van der Spek
Eindhoven University of Technology

It is assumed that serious games influences learning in 2 ways, by changing cognitive processes and by affecting motivation. However, until now research has shown little evidence for these assumptions. We used meta-analytic techniques to investigate whether serious games are more effective in terms of learning and more motivating than conventional instruction methods (learning: $k = 77$, $N = 5,547$; motivation: $k = 31$, $N = 2,216$). Consistent with our hypotheses, serious games were found to be more effective in terms of learning ($d = 0.29$, $p < .01$) and retention ($d = 0.36$, $p < .01$), but they were not more motivating ($d = 0.26$, $p > .05$) than conventional instruction methods. Additional moderator analyses on the learning effects revealed that learners in serious games learned more, relative to those taught with conventional instruction methods, when the game was supplemented with other instruction methods, when multiple training sessions were involved, and when players worked in groups.

Keywords: serious games, game-based learning, cognition, motivation, meta-analysis

In the last decade, researchers have propagated the use of computer games for the purpose of learning and instruction (often referred to as serious games or game-based learning). In this respect, serious games are hypothesized to address both the cognitive and the affective dimensions of learning (O'Neil, Wainess, & Baker, 2005), to enable learners to adapt learning to their cognitive needs and interests, and to provide motivation for learning (Malone, 1981). Reviews regarding the effects of serious games show ambiguous results (Ke, 2009; Sitzmann, 2011; Vogel et al., 2006; Wouters, van der Spek, & van Oostendorp, 2009), but several scholars have noted that in general the quality of game research is poor (O'Neil et al., 2005) and that serious games are not more effective in terms of learning than other instruction methods when they are tested scientifically (Clark, Yates, Early, & Moulton, 2010). Many claims are supported by anecdotal arguments and lack sound empirical evidence. However, in the last 5 years, more well-designed empirical studies investigating the ef-

fects of serious games on learning and motivation have been published.

Our goal in this study was to statistically summarize the research on the effects of serious games on learning and motivation. Mayer (2011) has divided game research into three categories: a value-added approach, which questions how specific game features foster learning and motivation; a cognitive consequences approach, which investigates what people learn from serious games; and a media comparison approach, which investigates whether people learn better from serious games than from conventional media. Our meta-analysis adopted the media comparison approach. We compared serious games with conventional instruction methods such as lectures, reading, drill and practice, or hypertext learning environments. In addition, this study discerned instructional and contextual factors that may moderate the effectiveness and motivational appeal of serious games. Several meta-analyses have been conducted with respect to the effects of serious games (Ke, 2009; Sitzmann, 2011; Vogel et al., 2006). The meta-analysis by Ke (2009) is an interesting exploration of the field of game-based learning, but it does not statistically summarize effect sizes. The Vogel et al. (2006) meta-analysis investigated both cognitive and attitudinal effects and found that computer games and interactive simulations yielded higher cognitive outcomes than did conventional learning methods. Our meta-analysis expanded this research by incorporating the high number of well-designed studies that have been published in recent years and by focusing on other instructional and contextual factors, such as the number of training sessions with serious games and the moment of measurement of the learning effects (immediate or delayed). The more recent meta-analysis by Sitzmann (2011) focuses on simulation games, whereas our research has a broader perspective on serious games. Although this study shares some moderator variables with

This article was published Online First February 4, 2013.

Pieter Wouters, Christof van Nimwegen, and Herre van Oostendorp, Institute of Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands; Erik D. van der Spek, Department of Industrial Design, Eindhoven University of Technology, Eindhoven, the Netherlands.

This research was funded by the Netherlands Organization for Scientific Research (Project No. 411-10-902). This research was also supported by the GATE project, funded by the Netherlands Organization for Scientific Research and the Netherlands ICT Research and Innovation Authority (ICT Regie).

Correspondence concerning this article should be addressed to Pieter Wouters, Utrecht University, Institute of Information and Computing Sciences, P.O. Box 80.089, 3508 TB Utrecht, the Netherlands. E-mail: P.J.M.Wouters@uu.nl

the Sitzmann study, we introduce new variables such as the domain in which the serious game is used, the age of the learners, and the group size (individual vs. group).

In the following sections we first define serious games. Next, we describe the theoretical framework with the main hypotheses and the moderator variables. The Method section comprises a description of the literature research, the inclusion criteria, the coding of the moderator variables, and the calculation of effect sizes. The Results section presents the general characteristics of the analysis, the main effects, and the effects of the moderator variables. Finally, we discuss the findings, draw conclusions, and depict some avenues for future research.

Definition of Serious Games

Several scholars have provided definitions or classifications of computer games characteristics (Garris, Ahlers, & Driskell, 2002; Malone, 1981; Prensky, 2001). For the purpose of this meta-analysis, we describe computer games in terms of being *interactive* (Prensky, 2001; Vogel et al., 2006), based on a set of *agreed rules and constraints* (Garris et al., 2002), and directed toward a clear *goal* that is often set by a *challenge* (Malone, 1981). In addition, games constantly provide *feedback*, either as a score or as changes in the game world, to enable players to monitor their progress toward the goal (Prensky, 2001). Some scholars contend that computer games also involve a competitive activity (against the computer, another player, or oneself), but it can be questioned if this is essentially a defining characteristic. Of course, there are many games in which the player is in competition with another player or with the computer, but in a game such as SimCity, players may actually enjoy the creation of a prosperous city that satisfies their beliefs or ideas without having the notion that they engage in a competitive activity. In the same vein, a narrative or the development of a story can be very important in a computer game (e.g., in adventure games), but again it is not a prerequisite for being a computer game (e.g., action games do not really require a narrative). In speaking of a serious (computer) game, we mean that the objective of the computer game is not to entertain the player, which would be an added value, but to use the entertaining quality for training, education, health, public policy, and strategic communication objectives (Zyda, 2005).

Theoretical Framework

In theory, games may influence learning in two ways, by changing the cognitive processes and by affecting the motivation. The (inter)active nature of computer game aligns with the current emphasis in educational psychology that active cognitive processing of educational material is a prerequisite for effective and sustainable learning (cf. Wouters, Paas, & van Merriënboer, 2008). Second, it is possible with computer games to simulate tasks in such a way that performing them in the game involves the same cognitive processes that are required for task performance in the real world (Tobias, Fletcher, Dai, & Wind, 2011). Finally, the immediate feedback in computer games provides players information regarding the correctness of their actions and decisions and thus gives them the opportunity to correct inaccurate information (Cameron & Dwyer, 2005; Moreno & Mayer, 2005).

Several classifications have been proposed for learning outcomes (for an overview, see Kraiger, Ford, & Salas, 1993; Wout-

ers et al., 2009). In this meta-analysis, we focus on the cognitive dimension of learning. In the Wouters et al. (2009) classification, this dimension is divided into knowledge and cognitive skills. Knowledge refers to encoded knowledge reflecting both text-oriented learning (e.g., verbal knowledge) and non-text-oriented learning (e.g., knowledge derived from an image). A cognitive skill pertains to more complex cognitive processes, such as in problem solving when a learner applies knowledge and rules to achieve a solution for a (novel) situation. With reference to the aforementioned arguments, we will make a distinction in learning between knowledge and cognitive skills.

Our first hypothesis contends

1. That instruction with serious games yields higher learning gains than conventional instruction methods.

In the majority of studies, learning is measured immediately after the learning stage. The question can be raised whether such an immediate test is appropriate when the focus is on sustainable learning, which occurs when learners are still able to adequately apply the learned knowledge and skills in the long term. It is still an exception to include a delayed test in experimental designs. In this meta-analysis, sufficient pairwise comparisons were available to justify the inclusion of retention as a learning variable. For simulation games, there is some support that the acquired knowledge and skills are maintained over time (Pierfy, 1977; Sitzmann, 2011; van der Spek, 2011). In line with these results, we expect

2. That instruction with serious games will yield a higher level of retention than conventional instructional methods.

Several theories emphasize the potential of serious games to positively influence intrinsic motivation (Garris et al., 2002; Malone, 1981). This means that players are willing to invest more time and energy in game play not because of extrinsic rewards but because the game play in itself is rewarding. Several characteristics of serious games have been identified for this motivating appeal. Malone (1981) proposed that the most important factors that make playing a computer game intrinsically motivating are challenge, curiosity, and fantasy. Two other essential factors associated with computer games, autonomy (i.e., the opportunity to make choices) and competence (i.e., a task is experienced as challenging but not too difficult), originate from self-determination theory and are known to positively influence the experienced motivation (Przybylski, Rigby, & Ryan, 2010; Ryan, Rigby, & Przybylski, 2006). We therefore hypothesize

3. That instruction with serious games is more motivating than conventional instruction methods.

Moderator Variables for Learning

We also investigate how situational and contextual variables may moderate learning with serious games. We distinguish between hypothesized moderators, nonhypothesized moderators, and methodological moderators. First, we describe and ground the hypothesized moderators.

Learning arrangement of the comparison group. Modern educational theories advocate active cognitive processing as a prereq-

quisite for genuine learning (Chi, de Leeuw, Chiu, & LaVancher, 1994; Mayer, 2001; Wittrock, 1974). In observational learning, for example, stronger learning effects are reported when learners engage in active coding (Bandura, 1976). Also, the research literature on self-explanations indicates that an active engagement of learners in the learning process fosters a better integration of new knowledge with prior knowledge and higher levels of transfer (Chi et al., 1994; Renkl & Atkinson, 2002; Roy & Chi, 2005). In this respect, a learning environment that stimulates an active cognitive attitude of the learners (e.g., doing practices and exercises) may foster more effective learning than does an arrangement in which learners are not explicitly prompted to actively engage in learning (e.g., reading an expository text or following a lecture). Therefore, the treatment that the comparison group receives may be an important moderator. We hypothesize that

4. The beneficial effect of serious games on learning is larger when the comparison group receives passive instruction than when the comparison group receives active instruction.

Serious game combined with other instructional methods.

In computer games, players typically act and see the outcome of their actions reflected in changes in the game world. This may lead to a kind of intuitive learning: They know how to apply knowledge, but they cannot explicate it (Leemkuil & de Jong, 2011). Yet, it is important that learners verbalize their knowledge, because it enables them to integrate new knowledge with their prior knowledge, resulting in better recall and higher transfer of learning (Wouters et al., 2008). It is possible that supplemental instruction methods (e.g., discussion, explicit practice) enable learners to engage in learning activities that further support the articulation of knowledge.

Some evidence comes from Sitzmann (2011), who found that arrangements in which a simulation game was supplemented with other instructional methods yielded higher levels of learning. In arrangements in which only a simulation game was used, the comparison group performed better. In line with this observation, we hypothesize that

5. Relative to the comparison group, learning arrangements in which serious games are supplemented with other instruction methods will yield higher learning gains than will arrangements in which serious games are the only instruction method.

Number of training sessions. The question can be raised whether a training of only one session is sufficient to ensure cognitive changes. Serious games can be complex learning environments for several reasons. For example, players may have to attend to different locations on the screen and coordinate this with mouse or joystick movements, or they may have to engage in a task in which multiple variables that mutually interact play a role. It is plausible that, in comparison to that of conventional instruction methods, the effectiveness of serious games in terms of learning pays off only after multiple training sessions in which the players get used to the game. We hypothesize that

6. Multiple training sessions with serious games will yield higher learning gains than will multiple training sessions with conventional instruction methods.

Group size. One argument for collaborative learning in computer games is that it supports learners in articulating the knowledge that would otherwise have remained intuitive (van der Meij, Albers, & Leemkuil, 2011), but research comparing collaborative and solitary game play is ambiguous. The observation by Inkpen, Booth, Klawe, and Uptis (1995) that collaborative play resulted in significantly higher scores on motivation and learning outcomes than did solitary play was not confirmed by van der Meij et al. (2011). The meta-analysis by Vogel et al. (2006) revealed that both single users and groups showed higher cognitive gains in interactive simulations and games than in conventional teaching methods, but the effect size for single users was much larger than for groups. On the basis of these observations, we hypothesize that

7. Compared with the comparison group, single users will yield higher learning gains than will players who play in a group.

In addition to investigating these hypothesis-oriented moderator variables, we investigated other variables that potentially may have a moderating effect on learning. These include instructional domain, age of the player, the level of realism, and the use of a narrative.

Instructional domain. Serious games are used in different domains, ranging from domains that are part of school curricula (e.g., biology, mathematics), job-oriented domains (e.g., military), to more basic cognitive processing (e.g., visual attention). Some domains may be more connected with learning with serious games than are other domains.

Age. The question can be raised whether age is a moderator. The meta-analysis by Vogel et al. (2006), however, did not find differences between age groups in learning with serious games. In the light of the large number of studies over the last 5 years, we addressed this question again.

Level of realism. Designers of serious games have neither the money nor the time to create computer games that can match commercial computer games in level of realism. It is sometimes argued that players have expectations about the design of serious games that are based on their experience with commercial computer games. In that case, it is not unlikely that they will become disappointed, which may be reflected in less motivation and learning. Vogel et al. (2006) investigated the level of realism (photo-realistic, high-quality cartoons, low-quality pictures, or unrealistic) in their meta-analysis but found no differences between the levels. The rapid technological developments and the increase in empirical studies in the last years justify a new examination of the impact of the level of realism.

Narrative. In game genres such as adventure games and role-playing games, narratives play an important role (Prensky, 2001). Research on learning from text shows that narratives foster learning and engagement. For example, compared with expository text or newspaper items, stories yield better recall, generate more inferences, and are more entertaining (Graesser, Singer, & Trabasso, 1994). Another argument for adding a narrative to the game is that it may scaffold problem solving during the game (Dickey, 2006). From a cognitive perspective, however, it can be argued that an engaging narrative may distract learners from the learning material and, given the limited cognitive capacity, withhold from them cognitive activities that yield learning (Mayer, Griffith, Naf-

taly, & Rothman, 2008). It is as yet unclear whether a narrative in a serious game will foster learning and engagement. Some value-added studies have been conducted, but they reported contradictory results. For example, McQuiggan, Rowe, Lee, and Lester (2008) found a negative effect of a narrative compared to a minimal narrative in a computer game, and Cordova and Lepper (1996) reported a beneficial effect when a narrative component (fantasy) was included.

Finally, we considered some methodological moderators. Meta-analyses allow a comparison of studies that use different experimental designs and different statistical methods. However, comparing studies with different degrees of methodological rigor may also obscure the results of the meta-analysis and thus jeopardize the conclusion that the weighted mean effect size is attributable to specific features of the studies and not to spurious factors that come with such studies. We identified three methodological indicators that potentially may influence the weighted mean effect size and the impact of study features.

Publication source. A potential danger in a meta-analysis is the “file drawer problem”: the concern that the studies in the meta-analysis are not a correct reflection of all studies that are actually conducted (Ellis, 2010) because studies published in peer-reviewed journals and/or proceedings are more likely to have achieved statistical significance and larger effect sizes than are studies that have not been published (Rosenthal, 1995). Therefore, we used the publication source (peer-reviewed journal, proceedings, and unpublished) as a moderator variable.

Randomization. Second, we took into account whether a pure or a quasi-experimental design was used. In the latter case, participants are not randomized between the conditions, which may allow alternative explanations for the results that are found.

Experimental design. Finally, we considered whether a posttest-only design or a pretest–posttest design was used.

Moderator Variables for Motivation

For motivation, the same moderator variables were used, but we did not formulate hypotheses. We used the moderators to explore whether and to what extent contextual and situational factors have an impact on the motivational appeal of serious games.

Method

Literature Search

We started with computer-based searches via Google Scholar. The search terms we used were *game-based learning*, *PC games*, *video game*, *computer video game*, *serious games*, *educational games*, *simulation games*, *virtual environments*, and *muve*. If necessary, these search terms were combined with *learning*, *instruction*, *training*, *motivation*, and *engagement*. In addition, we investigated the references of previous meta-analyses and reviews on the effectiveness of serious games (Fletcher & Tobias, 2006; Ke, 2009; O’Neil et al., 2005; Sitzmann, 2011; Vogel et al., 2006; Wouters et al., 2009). In order to find unpublished but relevant studies, we asked researchers and educators within our network of scholars whether they were aware of relevant studies for the meta-analysis. Our meta-analysis covered the period from 1990 to

2012. Our research located 190 studies, of which 38 studies met our inclusion criteria (see the next section).

Inclusion Criteria and Coding

There were four inclusion criteria. First, the experimental group learned the content of the domain through a serious game, either as the sole instruction method or in combination with other instructional methods. In addition, there was a comparison group that engaged in an alternative instructional method. Second, the serious games and comparison groups had to receive the same learning content. Third, the study reported data or indications that allowed us to calculate or estimate effect sizes (group means and standard deviations, *t* test, *F* test, etc.). Fourth, we focused on nondisabled participants. The characteristics of each study that, in addition to the effect sizes and the sample sizes, were coded are described next.

Learning and retention. Two categories of learning outcomes were used to classify learning. “Knowledge” was used when a test involved knowledge of concepts, principles, definitions, symbols, or facts (e.g., Papastergiou, 2009, on computer knowledge). Studies in which learners had to solve problems, make decisions, or apply rules to a situation were coded as “Cognitive skills” (e.g., Kebritchi, Hirumi, & Bai, 2010). Retention was coded when a delayed measure for learning was available (the low number of pairwise comparisons does not allow a further breakdown in knowledge and cognitive skills). In the majority of the studies the delayed test took place 1 to 5 weeks after the intervention, but in one study the delayed test took place after 27 weeks (Segers & Verhoeven, 2003).

Motivation. We adopted a broad view on motivation. In the majority of the studies, a questionnaire or survey was used to measure motivation (e.g., Parchman, Ellis, Christinaz, & Vogel, 2000), interest (e.g., Ritterfeld, Shen, Wang, Nocera, & Wong, 2009), engagement (e.g., van Dijk, 2010), or attitude toward the topic involved in the experiment (e.g., Miller & Robertson, 2010). In one study, ratings of observed engagement (Brom, Preuss, & Klement, 2011) were used as a measure for motivation.

Learning arrangement of the comparison group. “Active instruction” refers to instruction methods that explicitly prompt learners to learning activities (e.g., exercises, hypertext training). We also coded whether the focus of the activity was drill-and-practice oriented or problem-solving oriented. “Passive instruction” includes listening to lectures; receiving classical instruction; and reading textbooks, expository text, or a PowerPoint presentation. Studies in which a combination of active and passive instruction was used were coded as “Mixed instruction.” For example, Squire, Barnett, Grant, and Higginbotham (2004) used a comparison group with guided discovery involving interactive lectures, experiments, and demonstrations.

Serious game combined with other instructional methods. A study was coded “Inclusive” when the serious game was combined with other instructional methods (e.g., Kebritchi et al., 2010). When the serious game was the only instructional method, it was coded as “Exclusive” (e.g., Adams, Mayer, MacNamara, Koenig, & Wainess, 2012).

Number of training sessions. Studies in which learners engaged in only one training session with the serious game were coded “1 session.” The time of this session ranged from 18 min

(van Dijk, 2010) to 3 hr (Jong, Shang, Lee, Lee, & Law, 2006). Studies involving more than one session were coded as "Multiple sessions." The number of sessions varied from three (Gremmen & Potters, 1997) to 40 (Miller & Robertson, 2010).

Group size. When learners worked alone with the serious game, the study was coded "Individual." In the case of dyads or a larger group, the study was coded "Group." One study used a mix of individual and group game play (Kebritchi et al., 2010), but it was coded as Group.

Instructional domain. The studies in our research covered a broad range of domains. The domains biology, mathematics, language, and engineering were coded as such. Domains that were mentioned only a few times (aviation, computer science, geography, physics, military, and triage) were coded as "Other."

Age. The following categories were used: "Children" (until 12 years), "Preparatory education" (13 to 17 years), "Students" (18–24 years), and "Adults" (older than 25 years).

Level of realism. Games that were either textual (e.g., Gremmen & Potters, 1997) or schematic (e.g., mazelike type of games; see Papastergiou, 2009) were coded "Schematic." Cartoonlike games were coded as "Cartoon" (e.g., Brom et al., 2011), likewise photorealistic games were coded as "Realistic" (e.g., Kebritchi et al., 2010). In some studies, different types of serious games were used (e.g., Segers & Verhoeven, 2003). When the level of realism could be determined for all types, the study was coded under one of the aforementioned classifications (if all game types had the same level of realism) or as "Mixed" (if the game types had different levels of realism). "Unknown" was used when the level of realism of a game could not be determined.

Narrative. Games with a basic storyline (e.g., Ke, 2008) or more elaborated storyline (e.g., Barab et al., 2009) were coded as "Narrative." Games without a storyline were coded "Nonnarrative" (e.g., Cameron & Dwyer, 2005).

Methodological variables. For each study, three methodological variables were coded. To start with, the publication source could be a peer-reviewed journal, proceedings, or unpublished. Second, we coded whether or not participants were assigned randomly to the conditions. If schools or classes were randomly assigned to conditions and the learning effects were reported on an individual level, we coded the study as not random (e.g., Miller & Robertson, 2010). Finally, the experimental design of the study—either posttest only or pretest–posttest—was coded.

A random selection of 20 studies was coded independently by two raters. The mean intercoder agreement was 90.8%. Differences in codings were discussed until agreement was reached. The remaining studies were coded by the first author.

Calculation and Analysis of the Effect Sizes

Cohen's d was used as indicator for the effect size: The difference on the dependent variables (learning, retention, or motivation) between the serious game group and the comparison group was calculated and divided by the pooled standard deviation. When the means and/or standard deviations were not available, formulas were used to estimate the effect size based on data of the t test or the univariate F test (Glass, McGaw, & Smith, 1981), adjusted means, or gain scores (Hedges & Olkin, 1985).

Effect sizes for studies with small sample bias were corrected (cf. Hedges & Olkin, 1985). When multiple measurements were

used for learning, retention, or motivation, an average was calculated. It was subsequently used to estimate the effect size. When multiple learning outcomes and/or multiple treatment or comparisons groups were used, each pairwise combination of a learning outcome and/or treatment or comparison group was treated as an independent study. The sample size was adjusted to avoid the overrepresentation of studies with multiple pairwise comparisons. For this purpose, we developed a procedure to assure that no comparison received an inappropriate weight (see the Appendix for a description of the procedure and an example).

We used the random-effects model for the main analyses and the moderator analyses with 95% confidence intervals around the weighted mean effect sizes. To calculate the effect sizes, we created a program in Excel using the formulas provided by Ellis (2010) and Borenstein, Hedges, Higgins, and Rothstein (2009).

Results

In total, 39 studies were identified; they yielded 77 pairwise comparisons on learning outcomes, 17 pairwise comparisons on retention, and 31 pairwise comparisons on motivation. Although we focused on studies conducted after 1990, 54% of the studies were conducted in the last 5 years (2007–2012). In total, 5,547 participants were involved. The sample sizes of the studies ranged from 16 to 1,105 participants. Table 1 (learning and retention) and Table 2 (motivation) present all included pairwise comparisons with effect sizes and their classification on the nonmethodological moderator variables.

The heterogeneity of effect sizes was confirmed only for learning ($Q_{total} = 323.79$, $df = 76$, $p < .001$) and for motivation ($Q_{total} = 71.05$, $df = 30$, $p < .001$) but not for retention ($Q_{total} = 8.68$, $df = 16$, $p > .05$). Therefore, a moderator analysis is justified for learning and motivation. For all analyses, alpha was set at .05.

Main Effect Analysis

The weighted mean effect sizes are presented in Table 3. Although we included a methodological moderator to examine a possible publication bias, we also calculated the fail-safe N , which is the number of studies averaging null results that has to be retrieved in order to reject the summary effect size. A publication bias is unlikely to occur when the fail-safe N (for this study, 3,489) exceeds the suggested threshold of the quintuple of pairwise comparisons plus 10 (Ellis, 2010), which is clearly the case in this review: $3,489 > 5 \times 77 + 10 = 395$.

The first hypothesis, which predicts that instruction with serious games yields higher learning gains than conventional instruction, is confirmed. The weighted mean effect size of 0.29 for learning in favor of serious games is statistically significant ($z = 4.67$, $p < .001$). Also, the effect sizes of knowledge and cognitive skills show that serious games are superior to conventional instructional methods (knowledge: $d = 0.27$, $z = 2.00$, $p < .05$; cognitive skills: $d = 0.29$, $z = 4.12$, $p < .001$). The comparisons of the effect sizes of both learning outcomes reveal no significant differences ($p > .1$). We also tested the homogeneity of effect sizes for the two learning outcomes. The Q_b statistic, $\chi^2(1) = 3.86$, $p > .1$, suggests that the differences between the two learning outcomes are attributable to sampling error. For this reason we used the overall learning effect size ($d = 0.29$) in the subsequent moderator analysis.

Table 1
Studies and Pairwise Comparisons of Serious Games vs. a Comparison Group and the Effects on Learning

Study	Adjusted <i>N</i>	Learning outcome	<i>d</i> _{immediate}	<i>d</i> _{retention}	Activity comparison group	Inclusive/exclusive	No. sessions	Group size	Domain	Age	Level of realism
Adams et al. (2012), Experiment 1	21	Knowledge	-1.34		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Realistic
Adams et al. (2012), Experiment 2	21	Skills	-0.36		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Realistic
Adams et al. (2012), Experiment 2 Narrative	57	Skills	-0.31		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Realistic
Adams et al. (2012), Experiment 2 Nonnarrative	57	Skills	-0.47		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Realistic
Annetta et al. (2009)	130	Skills	0.00		Active (practice)	Inclusive	1	Group	Biology	Preparatory education	Realistic
Bai et al. (2012)	219	Skills	0.19		Mixed	Inclusive	>1	Group	Math	Preparatory education	Realistic
Barab et al. (2009)											
Pilot study	35	Skills	0.66		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Realistic
Pilot study	35	Skills	0.87		Active (hypertext)	Exclusive	1	Individual	Biology	Students	Realistic
Main study	12	Skills	0.76		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Realistic
Main study	12	Skills	1.39		Passive (PowerPoint)	Exclusive	1	Group	Biology	Students	Realistic
Main study	12	Skills	0.29		Mixed	Exclusive	1	Individual	Biology	Students	Realistic
Main study	12	Skills	0.69		Mixed	Exclusive	1	Group	Biology	Students	Realistic
Main study	65	Skills	0.47		Active (assignments)	Inclusive	>1	Individual	Biology	Preparatory education	Realistic
Barab et al. (2012)	54	Skills	0.28		Active (Internet task)	Exclusive	1	Individual	Other	Students	Realistic
Barlett et al. (2009)	59	Skills	0.19		Active (Internet task)	Exclusive	1	Individual	Other	Students	Realistic
Betz (1996)	24	Skills	- ^a	0.64	Passive (reading)	Inclusive	>1	Individual	Other	Students	Cartoon
Brom et al. (2011)	100	Knowledge	0.03	0.25	Passive (lecture)	Inclusive	1	Individual	Biology	Preparatory education	Cartoon
Cameron & Dwyer (2005)											
Game	140	Skills	0.21		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Schematic
Game with elaborated feedback	144	Skills	0.50		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Schematic
Game with responsive feedback	139	Skills	0.79		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Students	Schematic
Dede et al. (2005)	125	Knowledge	0.23		Mixed	Inclusive	>1	Individual	Biology	Preparatory education	Realistic
Gremmen & Potters (1997)	38	Skills	0.59	0.67	Passive (lecture)	Exclusive	>1	Group	Other	Students	Schematic
Jarvis & de Freitas (2009)	91	Skills	0.67		Active (practice)	Inclusive	1	Individual	Other	Adults	Realistic
Jong et al. (2006)	158	Skills	-0.09		Active (hypertext)	Inclusive	1	Individual	Math	Preparatory education	Cartoon
Ke (2008)	96	Skills	0.24		Active (practice)	Exclusive	>1	Mix	Math	Preparatory education	Cartoon
Ke & Grabowski (2007)											
Competitive	59	Skills	0.32		Active (practice)	Exclusive	>1	Individual	Math	Preparatory education	Cartoon
Cooperative	61	Skills	0.39		Active (practice)	Exclusive	>1	Group	Math	Preparatory education	Cartoon
Kebrichi et al. (2010)	193	Skills	0.27		Mixed	Inclusive	>1	Group	Math	Preparatory education	Realistic
Laffey et al. (2003)	56	Skills	0.96		Mixed	Inclusive	>1	Individual	Math	Children	Cartoon

(table continues)

Table 1 (continued)

Study	Adjusted <i>N</i>	Learning outcome	$d_{\text{immediate}}$	$d_{\text{retention}}$	Activity comparison group	Inclusive/exclusive sessions	No. sessions	Group size	Domain	Age	Level of realism
McQuiggan et al. (2008) Narrative	77	Skills	-0.98		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Preparatory education	Realistic
Nonnarrative	73	Skills	-0.61		Passive (PowerPoint)	Exclusive	1	Individual	Biology	Preparatory education	Realistic
Miller & Robertson (2010)	42	Skills	0.46		Active (brain gym)	Inclusive	>1	Individual	Math	Children	Unknown
Miller & Robertson (2011)	635	Skills	0.08		Mixed	Inclusive	>1	Individual	Math	Children	Unknown
Moreno et al. (2001) Experiment 3	19	Knowledge	0.85		Active (multimedia)	Exclusive	1	Individual	Biology	Students	Schematic
Experiment 3	19	Skills	0.60		Active (multimedia)	Exclusive	1	Individual	Biology	Students	Schematic
Nicol & Anderson (2000)	12	Skills	-0.06	0.04	Active (exercises)	Exclusive	>1	Group	Math	Adults	Unknown
Okolo (1992)	41	Skills	0.00		Active (practice)	Exclusive	>1	Individual	Math	Students	Unknown
Papastergiou (2009)	88	Knowledge	0.61		Mixed	Inclusive	1	Individual	Other	Preparatory education	Schematic
Parchman et al. (2000) CBDP	10	Knowledge ^b	-0.91		Active (practice)	Exclusive	1	Individual	Engineering	Students	Cartoon
CBDP	10	Skills	0.32		Active (practice)	Exclusive	1	Individual	Engineering	Students	Cartoon
ECBI	15	Knowledge	-0.57		Mixed	Exclusive	1	Individual	Engineering	Students	Cartoon
ECBI	15	Skills	-0.47		Mixed	Exclusive	1	Individual	Engineering	Students	Cartoon
CI	15	Knowledge	0.21		Mixed	Exclusive	1	Individual	Engineering	Students	Cartoon
CI	15	Skills	-0.31		Mixed	Exclusive	1	Individual	Engineering	Students	Cartoon
Ricci et al. (1996)	30	Knowledge	-0.27	0.32	Active (test)	Exclusive	1	Individual	Other	Students	Schematic
	30	Knowledge	1.28	0.47	Passive (reading)	Exclusive	1	Individual	Other	Students	Schematic
Ritterfeld et al. (2009)	19	Knowledge	0.32	0.46	Active (hypertext)	Exclusive	1	Individual	Biology	Students	Realistic
	19	Skills	0.25	0.33	Active (hypertext)	Exclusive	1	Individual	Biology	Students	Realistic
	19	Knowledge	0.62	0.68	Passive (reading)	Exclusive	1	Individual	Biology	Students	Realistic
	19	Skills	0.28	0.15	Passive (reading)	Exclusive	1	Individual	Biology	Students	Realistic
Rosas et al. (2003) Reading	368	Skills	0.02		Mixed	Inclusive	>1	Individual	Language	Preparatory education	Cartoon
Spelling	368	Skills	-0.10		Mixed	Inclusive	>1	Individual	Language	Preparatory education	Cartoon
Mathematics	368	Skills	0.08		Mixed	Inclusive	>1	Individual	Math	Preparatory education	Cartoon
Seelhammer (2009)	20	Knowledge	-1.60		Passive (reading)	Inclusive	1	Individual	Biology	Preparatory education	Realistic
	20	Skills	-0.57		Passive (reading)	Inclusive	1	Individual	Biology	Preparatory education	Realistic
Segers & Verhoeven (2003) Immigrants, year 1	33	Knowledge	0.10	0.01	Mixed	Inclusive	>1	Individual	Language	Children	Cartoon
Immigrants, year 2	62	Knowledge	0.34	0.36	Mixed	Inclusive	>1	Individual	Language	Children	Cartoon
Natives, year 1	34	Knowledge	0.23	0.10	Mixed	Inclusive	>1	Individual	Language	Children	Cartoon
Natives, year 2	18	Knowledge	0.33	0.26	Mixed	Inclusive	>1	Individual	Language	Children	Cartoon
Sindre et al. (2009)	32	Knowledge	0.33		Active (exercises)	Inclusive	1	Individual	Other	Students	Unknown
	32	Knowledge	0.38		Passive (reading)	Inclusive	1	Individual	Other	Students	Unknown
	90	Skills	1.75		Mixed	Inclusive	>1	Individual	Other	Preparatory education	Cartoon
Squire et al. (2004) Suh et al. (2010) Listening	55	Skills	2.22		Mixed	Inclusive	>1	Group	Language	Preparatory education	Unknown
Speaking	55	Skills	0.22		Mixed	Inclusive	>1	Group	Language	Preparatory education	Unknown
Reading	55	Skills	1.20		Mixed	Inclusive	>1	Group	Language	Preparatory education	Unknown
Writing	55	Skills	1.71		Mixed	Inclusive	>1	Group	Language	Preparatory education	Unknown
van Dijk (2010)	24	Knowledge	-0.78	1.07	Passive	Exclusive	1	Individual	Other	Students	Realistic
					(PowerPoint)						
	24	Knowledge	-0.33	1.06	Passive	Exclusive	1	Individual	Other	Students	Realistic
					(PowerPoint)						

(table continues)

Table 1 (continued)

Study	Adjusted <i>N</i>	Learning outcome	<i>d</i> _{immediate}	<i>d</i> _{retention}	Activity comparison group	Inclusive/exclusive sessions	No. sessions	Group size	Domain	Age	Level of realism
Van Eck & Dempsey (2002) Adventure/computer	23	Skills	0.08		Active (word problem)	Exclusive	1	Individual	Math	Preparatory education	Schematic
No adventure/computer	24	Skills	0.45		Active (word problem)	Exclusive	1	Individual	Math	Preparatory education	Schematic
Adventure/no computer	23	Skills	0.52		Active (word problem)	Exclusive	1	Individual	Math	Preparatory education	Schematic
No adventure/no computer	18	Skills	-0.28		Active (word problem)	Exclusive	1	Individual	Math	Preparatory education	Schematic
Virvou et al. (2005) Part 1	90	Knowledge	0.96		Mixed	Exclusive	>1	Individual	Other	Preparatory education	Realistic
Part 2, poor performance	30	Knowledge	1.79		Mixed	Exclusive	>1	Individual	Other	Preparatory education	Realistic
Part 2, medium performance	30	Knowledge	0.84		Mixed	Exclusive	>1	Individual	Other	Preparatory education	Realistic
Part 2, good performance	30	Knowledge	0.10		Mixed	Exclusive	>1	Individual	Other	Preparatory education	Realistic
Wrzesien (2010)	48	Knowledge	0.28		Passive (reading)	Exclusive	1	Group	Biology	Children	Realistic
Yip & Kwan (2006)	100	Knowledge	1.20		Active	Inclusive	>1	Individual	Language	Students	Unknown

Note. CDBP = computer-based drill and practice; ECBI = experimental computer-based instruction; CI = classical instruction.

^a In the Betz (1996) study, learning was measured after a week. It was therefore classified as retention. ^b In the Parchman et al. (2000) study, knowledge of definitions and symbols was defined as knowledge and principle and rule application was defined as cognitive skills. The column Activity comparison group describes the instruction activity and how it was classified in the meta-analysis.

The second hypothesis predicts that instruction with serious games yields a higher level of retention than training with conventional instructional methods. Indeed, the results show that the superiority of serious games over conventional instructional methods is maintained in a delayed test ($d = 0.36$, $z = 2.41$, $p < .05$).

The third hypothesis predicts that serious games are more motivating than conventional instructional methods. Although the summary effect size is 0.26 in favor of serious games, the corresponding z score indicates that the difference in motivation is not statistically significant ($z = 1.77$, $p > .05$).

Moderator Analysis

We conducted a moderator analysis only for learning and motivation. For retention, a moderator analysis was not appropriate, given the homogeneous distribution of effect sizes (see the aforementioned Q_{total} statistic) and the low number of pairwise comparisons.

In order to compare the subgroups in the moderator analysis, we adopted the z -testing method for random effects with separate estimates of between-study variance (see Borenstein et al., 2009). When a moderator variable comprised more than two categories, the Holm-Bonferroni procedure was used to adjust the critical p value to control for the Type 1 error (cf. Ginns, 2005). In Holm's sequential version, the results of the Bonferroni tests are ordered from the smallest to the highest p value. The test with the lowest p value is then tested first with a Bonferroni correction involving all tests. The second test applies a Bonferroni correction involving one test less. This procedure continues for the remaining tests (Abdi, 2010).

Moderator analysis for learning. The results of the moderator analysis for learning are shown in the left part of Table 4.

The fourth hypothesis predicts that serious games yield more learning when the comparison group engages in passive instruction rather than active instruction. This hypothesis is not confirmed ($z_{active-passive} = -1.38$, $p > .05$). On the contrary, serious games do not improve learning more than does passive instruction. The beneficial effect of serious games is larger for mixed instructional methods than for passive instruction methods ($z_{mixed-passive} = 2.56$, $p = .005$). All other comparisons revealed no significant differences ($ps > .05$). Although the effect of serious games seems stronger for instruction with a focus on problem solving ($d = 0.31$) than for drill-and-practice-oriented instruction ($d = 0.22$), the difference is not significant ($z_{drill-and-practice-problem\ solving} = 0.45$, $p > .1$).

In Hypothesis 5, we expect that, for the experimental relative to the comparison group, serious games supplemented with other instructional methods will yield higher learning gains than serious games without supplemental instructional methods. The results confirm this hypothesis: Compared with conventional instruction methods, serious games yield higher learning gains irrespective of whether they are presented alone ($d = 0.20$) or supplemented with other instructional methods ($d = 0.41$), but learners learn most when serious games are supplemented with other instructional methods ($z_{inclusive-exclusive} = 1.66$, $p = .048$).

The sixth hypothesis predicts that multiple training sessions with serious games will yield higher learning gains than multiple training sessions with conventional instruction methods. When only one training session is involved, serious games are not more

Table 2
Studies and Pairwise Comparisons of Serious Games vs. a Comparison Group and the Effects on Motivation

Study	Adjusted <i>N</i>	Motivation	<i>d</i>	Activity comparison group	Inclusive/exclusive	No. sessions	Group size	Domain	Age	Level of realism
Annetta et al. (2009)	72	Observed engagement	0.81	Active (practice)	Inclusive	1	Group	Biology	Preparatory education	Realistic
Bai et al. (2012)	219	ARCS	0.30	Mixed	Inclusive	>1	Group	Math	Preparatory education	Realistic
Barab et al. (2012)	33	Question engagement	1.59	Active (assignments)	Inclusive	>1	Individual	Biology	Preparatory education	Realistic
Brom et al. (2012)	50	Question appeal	0.02	Passive (lecture)	Inclusive	1	Individual	Biology	Preparatory education	Cartoon
Ke (2008)	50	Question value	-0.47	Passive (lecture)	Inclusive	1	Individual	Biology	Preparatory education	Cartoon
Ke & Grabowski (2007)	358	ATMI	0.47	Active (practice)	Exclusive	>1	Mix	Math	Preparatory education	Cartoon
Competitive										
Cooperative										
Kebritchi et al. (2010)	59	ATMI	0.18	Active (practice)	Exclusive	>1	Individual	Math	Preparatory education	Cartoon
	61	ATMI	0.55	Active (practice)	Exclusive	>1	Group	Math	Preparatory education	Cartoon
Kuo (2007)	193	Survey	-0.23	Mixed	Inclusive	>1	Group	Math	Preparatory education	Realistic
	23	Survey	0.67	Active (explore)	Exclusive	>1	Individual	Biology	Children	Cartoon
	23	Visiting times	0.76	Active (explore)	Exclusive	>1	Individual	Biology	Children	Cartoon
Miller & Robertson (2010)	14	Learning self-concept	0.46	Active (brain gym)	Inclusive	>1	Individual	Math	Children	Unknown
	14	Self-esteem	0.13	Active (brain gym)	Inclusive	>1	Individual	Math	Children	Unknown
	14	Math self-concept	0.05	Active (brain gym)	Inclusive	>1	Individual	Math	Children	Unknown
Miller & Robertson (2011)	212	Learning self-concept	0.00	Mixed	Inclusive	>1	Individual	Math	Children	Unknown
	212	Self-esteem	-0.08	Mixed	Inclusive	>1	Individual	Math	Children	Unknown
	212	Math self-concept	0.12	Mixed	Inclusive	>1	Individual	Math	Children	Unknown
Moreno et al. (2001)										
Experiment 3	19	Question motivation	0.10	Active (multimedia)	Exclusive	1	Individual	Biology	Students	Schematic
Experiment 3	19	Question interest	0.24	Active (multimedia)	Exclusive	1	Individual	Biology	Students	Schematic
Papastergiou (2009)	88	Question	0.41	Mixed	Inclusive	1	Individual	Other	Preparatory education	Schematic
Parchman et al. (2000)	20	ARCS	0.01	Active (practice)	Exclusive	1	Individual	Engineering	Students	Cartoon
	30	ARCS	-0.72	Mixed	Exclusive	1	Individual	Engineering	Students	Cartoon
	31	ARCS	-0.12	Active (practice)	Exclusive	1	Individual	Engineering	Students	Cartoon
Ricci et al. (1996)	30	Question	0.57	Active (tests)	Exclusive	1	Individual	Other	Students	Schematic
	30	Question	1.21	Passive (reading)	Exclusive	1	Individual	Other	Students	Schematic
Ritterfeld et al. (2009)	38	Question	0.37	Active (hypertext)	Exclusive	1	Individual	Biology	Students	Realistic
	38	Question	0.34	Passive (text)	Exclusive	1	Individual	Biology	Students	Realistic
van Dijk (2010)	24	Question	0.25	Passive (PowerPoint)	Exclusive	1	Individual	Medicine	Students	Realistic
Wrzesien et al. (2010)	24	Question engagement	0.66	Passive (reading)	Exclusive	1	Group	Biology	Children	Realistic
	24	Question enjoyment	0.89	Passive (reading)	Exclusive	1	Group	Biology	Children	Realistic
	24	Question motivation	0.00	Passive (reading)	Exclusive	1	Group	Biology	Children	Realistic

Note. The column Activity comparison group describes the instruction activity and how it was classified in the meta-analysis (between brackets). ATMI = Attitude Towards Mathematics Inventory; ARCS = Attention Relevance Confidence Satisfaction.

Table 3
Main Effects for Learning, Retention, and Motivation Comparing Serious Games With Other Instructional Methods

Variable	<i>d</i>	<i>SE</i>	<i>k</i>	<i>N</i>	95% CI	<i>Q_i</i>
Learning	0.29**	.06	77	5,547	[0.17, 0.42]	323.97
Knowledge	0.27*	.14	25	948	[0.01, 0.54]	90.12
Skills	0.29**	.07	52	4,599	[0.15, 0.43]	226.61
Retention	0.36*	.16	16	499	[0.07, 0.68]	8.68
Motivation	0.26	.15	31	2,216	[−0.03, 0.56]	71.05

Note. *d* = weighted mean effect size (* *p* > .05; ** *p* > .001); *SE* = standard error of the effect size; *k* = number of pairwise comparisons; *N* = sum of the sample sizes of each pairwise comparison; CI = confidence interval; *Q_i* = homogeneity statistic.

effective than conventional instruction methods. However, consistent with the hypothesis, the results also show that multiple sessions yield higher learning gains for serious games than for conventional instruction methods (*d* = 0.54). Additionally, the comparison of the groups reveal that multiple sessions are more effective than only one session ($z_{I\text{ session-multiple sessions}}$ = 3.94, *p* = .003).

In Hypothesis 7, we predict that, for the experimental relative to the comparison group, learners learn more when they individually play serious games than when they play in a group. Not only do the results reject the hypothesis, but they also show that the reverse is the case: With serious games, both learners playing individually and those playing in a group learn more than the comparison group (respectively, *d* = 0.22 and *d* = 0.66), but learners who play serious games in a group learn more ($z_{individual-group}$ = 2.34, *p* = .01).

In general, the results show that serious games improve learning more than conventional instruction methods in all domains except biology and engineering, but there is also much variation between the domains. Serious games are particularly effective in language (*d* = 0.66). For the experimental relative to the comparison group, serious games yield more learning in language than in biology ($z_{language-biology}$ = 2.28, *p* = .01) and mathematics ($z_{language-math}$ = 2.25, *p* < .01).

Serious games are superior to the comparison group for all age groups with the exception of adults. The comparisons of age groups reveal no differences (*ps* > .1). With respect to the level of realism, the results indicate that instruction with schematic serious games is superior to conventional instruction methods (*d* = 0.46). This is not true for cartoonlike or realistic serious games (*p* > .05). Mutual comparisons also show that schematic serious games are more effective than cartoonlike or realistic games ($z_{schematic-cartoonlike}$ = 1.89, *p* = .03, $z_{schematic-realistic}$ = 2.25, *p* = .01). Compared with conventional instruction methods, serious games without a narrative seem to be more effective than serious games with a narrative, but the difference is not significant ($z_{narrative-no\ narrative}$ = 1.34, *p* = .09).

Turning to the methodological moderators, we see that only studies in peer-reviewed journals report higher learning gains for serious games. For proceedings and unpublished papers the effect sizes are even negative, but it should be noted that the number of pairwise comparisons in both publication sources is very low. Comparisons based on the Holm–Bonferroni procedure show no significant differences between the publication sources (*ps* > .05).

The beneficial effect of serious games is contingent on the experimental rigor: Random assignment attenuates the effect of serious games ($z_{random-nonrandom}$ = 2.75, *p* < .003). In fact, in studies with randomization, serious games are not more effective than conventional instruction methods. Finally, the experimental design of the study (posttest only: *d* = 0.25 vs. pretest–posttest design: *d* = 0.32) does not have an impact on the magnitude of the effect size ($z_{posttest\ only-preposttest}$ = 0.55, *p* > .1).

Moderator analysis for motivation. The right side of Table 4 shows the moderator analysis for motivation. Two interesting observations can be made. First, serious games are more motivating compared to a group receiving active instruction (*d* = 0.45, *p* = .02). Second, relative to conventional instruction methods, serious games are more motivating when they are not combined with other instruction methods (*d* = 0.37, *p* = .03). We also found that, relative to a group receiving conventional instruction, schematic serious games are more motivating (*d* = 0.51, *p* = .02), but this conclusion is based on only five pairwise comparisons. All other moderators are not statistically significant. No comparisons of the subgroups within the moderator variables reached statistical significance (*ps* > .1).

Discussion

It is often argued that the affordances of computer games can be used to foster learning and motivation in instruction. Indeed several reviews have—at least partly—shown this potential (Ke, 2009; Vogel et al., 2006; Wouters et al., 2009), but the increase in empirical studies on serious games in the last 5 years justifies a new meta-analysis. In addition, it is still not clear which instructional and contextual factors have an impact on the effectiveness of serious games. Our results confirm the findings of the earlier reviews that, in general, serious games are more effective than conventional instruction methods. However, there are also some striking differences, which we discuss in this section.

Learning

The results on knowledge and cognitive skills suggest that training with serious games is more effective than training with conventional instruction methods. In line with Sitzmann (2011), the retention outcome shows that the cognitive gains are not attributable to the “freshness” of the learning material but that these gains persist in the long term. This retention effect is impor-

Table 4

Moderator Analysis Comparing Serious Games With Other Instructional Methods for Learning and Motivation

Variable	Learning			Motivation		
	<i>d</i>	<i>k</i>	95% CI for <i>d</i>	<i>d</i>	<i>k</i>	95% CI for <i>d</i>
Activity comparison group						
Active instruction	0.28***	24	[0.12, 0.45]	0.45*	13	[0.09, 0.80]
Drill and practice	0.22**	13	[0.08, 0.35]	0.27	9	[−0.20, 0.75]
Problem solving	0.31*	6	[0.01, 1.01]	0.88**	4	[0.31, 1.44]
Unknown	0.28*	5	[0.02, 0.54]			
Passive instruction	0.06	25	[−0.20, 0.33]	0.24	12	[−0.32, 0.81]
Mixed	0.50***	28	[0.30, 0.70]	0.07	6	[−0.62, 0.76]
Computer game alone						
Inclusive	0.41***	29	[0.23, 0.59]	0.18	13	[−0.33, 0.69]
Exclusive	0.20*	48	[0.03, 0.37]	0.37*	18	[0.07, 0.67]
No. training sessions						
One session	0.10	47	[−0.07, 0.26]	0.26	17	[−0.21, 0.73]
Multiple sessions	0.54***	30	[0.35, 0.72]	0.26	14	[−0.13, 0.65]
Group size ^a						
Individual	0.22**	63	[0.09–0.36]	0.23	24	[−0.14, 0.59]
Group	0.66**	13	[0.32, 1.00]	0.35	6	[−0.31, 1.01]
Domain						
Biology	0.11	28	[−0.11, 0.33]	0.44	13	[−0.03, 0.91]
Math/arithmetic	0.17**	16	[0.07, 0.28]	0.15	11	[−0.30, 0.60]
Language	0.66**	11	[0.25, 1.07]	—	—	—
Engineering	−0.36	6	[−0.80, 0.09]	0.24	7	[−0.62, 1.10]
Others	0.54***	16	[0.23, 0.85]	—	—	—
Age						
Children	0.30**	8	[0.08, 0.52]	0.15	11	[−0.30, 0.60]
Preparatory education	0.33**	31	[0.13, 0.54]	0.32	10	[−0.15, 0.79]
Students	0.23*	36	[0.04, 0.42]	0.22	10	[−0.49, 0.93]
Adults	0.50	2	[−0.10, 1.10]	—	—	—
Level of realism						
Schematic	0.46***	14	[0.27, 0.65]	0.51*	5	[0.05, 0.96]
Cartoonlike	0.20	20	[−0.01, 0.40]	0.12	15	[−0.42, 0.67]
Photorealistic	0.14	32	[−0.08, 0.35]	0.40	9	[−0.16, 0.97]
Mixed	—	—	—	—	—	—
Unknown	0.72**	11	[0.27, 1.16]	0.71	2	[0.01, 1.42]
Narrative						
Yes	0.25***	62	[0.11, 0.39]	0.32	17	[−0.08, 0.71]
No	0.46**	15	[0.18, 0.73]	0.19	14	[−0.27, 0.64]
Methodological variables						
Publication source						
Peer-reviewed journal	0.36***	67	[0.24, 0.48]	0.24	28	[−0.07, 0.56]
Proceedings	−0.16	7	[−0.91, 0.58]	—	—	—
Unpublished	−0.20	3	[−0.83, 0.43]	0.55	3	[−0.10, 1.20]
Randomization						
Yes	0.08	35	[−0.13, 0.29]	0.36	8	[−0.15, 0.86]
No	0.44***	42	[0.29, 0.60]	0.25	23	[−0.10, 0.60]
Design						
Posttest only	0.25**	27	[0.07, 0.44]	0.12	9	[−1.02, 1.25]
Pre-posttest	0.32***	50	[0.16, 0.48]	0.30	22	[0.01, 0.60]

Note. *d* = weighted mean effect size (* $p < .05$; ** $p < .01$; *** $p < .001$); *k* = number of pairwise comparisons; CI = confidence interval.

^a Ke (2008) provided combined data only for cooperative and individualistic conditions. Therefore, this study is not considered in the user group variable.

tant, because it supports what teachers and instructors deem important: that serious games lead to well-structured prior knowledge on which learners can build on during their learning career.

The meta-analysis also distinguishes some situational and contextual factors. The positive effect of multiple training sessions on learning is larger for serious games than for conventional instruction methods. We assumed that the advantages of serious games would emerge when the players engaged in more training sessions and became used to the complex learning environment. However, the results also allow other explanations. For example, with respect to text comprehension, Kintsch, Welsch, Schmalhofer, and Zimny

(1990) have shown that memory for the surface level and textbase-level representation of text decays over time, whereas memory for the situation model is robust to such decay. Perhaps immediately after learning from conventional instruction or the game, the textbase representation is still sufficiently available, causing no difference between the conventional instruction and game conditions. In contrast, after a decay of 2 to 4 days, students may need the situation model of the text to perform adequately on the test; then, the benefit by deeper processing in the game condition pays off (cf. Kintsch, 1998, p. 328). Some evidence for this assertion comes from the retention measure. Studies with a one-session

learning stage in which an immediate and a delayed test is administered show no efficacy on the short term ($k = 9$, $d = .14$, $p < .1$), but they do in the long term ($k = 9$, $d = .40$, $p < .01$). However, some caution is warranted. It is possible that in the short term such brief training session cause worse learning and less motivation than other instruction methods, whereas in the long term positive effects may appear. For example, players may voluntarily play the game without being asked and in this way learn. It is also possible that they actually have learned but that the type of test that is administered (e.g., a knowledge test asking definitions of concepts) does not detect the deep level of knowledge. In this respect, we propose to include other methods to measure learning (Day, Arthur, & Gettman, 2001; Wouters, van der Spek, & van Oostendorp, 2011).

Our hypothesis predicting that serious games are more effective when the comparison group engages in passive instruction rather than in active instruction is not confirmed. On the contrary, serious games are not more effective than passive instruction. These results seem to contradict those of Sitzmann (2011), who found that simulation games were far more effective when compared with passive instruction than with active instruction. A closer examination of the results shows that this moderator confounds with the number of instruction sessions moderator, because almost all studies involving passive instruction are conducted during a learning stage of one session. In that case, the failure to find a positive effect of serious games over passive instruction may be attributable to the one-session learning stage. This conclusion is supported by a similar pattern for active instruction (1 session: $k = 16$, $d = 0.18$; >1 session: $k = 8$, $d = 0.43$) and mixed instruction (1 session: $k = 7$, $d = 0.15$; >1 session: $k = 21$, $d = 0.58$).

Another significant moderator is whether serious games are used as the only instructional method or are supplemented with other instructional methods. The meta-analysis shows that serious games are more effective when they are supplemented with other instructional methods than they are when used as sole instruction method. This may be due to the fact that game players in the latter case gain intuitive knowledge, but they are not prompted to verbalize the new knowledge and so do not anchor it more profound in their knowledge base (Leemkuil & de Jong, 2011; Wouters et al., 2008). The additional effect of supplemental instructional methods is that they prompt or support players to articulate the new knowledge and integrate it with their prior knowledge. These findings are also in line with other research showing that the active reflection or reviewing of information and experiences is beneficial for learning. For example, regarding pure versus guided discovery learning research has shown that learning by doing has to be supplemented with opportunities to reflect (cf. Mayer, 2004). Likewise, in the game cycle model of Garris et al. (2002), debriefing, defined as the review and analysis of events that occurred in the game itself, is regarded as the most critical part of the (serious) game experience. This finding is also useful from a practical point of view. Practitioners such as teachers are still reluctant to adopt serious games in the classroom. One of the perceptions is that it is difficult to integrate the serious game in their daily practice (cf. Baek, 2008), but the results show the potential of using serious games together with instruction methods that they already use in the classroom.

Contrary to our hypothesis, serious games are more effective when played in groups (in most studies the participants played in

dyads) than when played alone. We proposed earlier that serious games foster some learning activities but that other learning activities, such as the articulation of knowledge, are not automatically addressed. These learning activities can be prompted by supplementing serious games with other instruction methods. The large effect of playing in a group suggests that this is also an effective method to incite these additional learning activities. However, this remains unclear, because most studies did not accurately describe the type of guidelines the players received for the collaboration. More research is needed, as well as a better understanding about the most effective group size (dyads, as in Annetta, Minogue, Holmes, & Chen, 2009, or many players, as in Suh, Kim, & Kim, 2010).

The results of the domain variable are difficult to interpret, because the variable confounds with other moderator variables. Remarkable is the large effect size for language. Rich multimodal environments such as computer games have characteristics that appear to be beneficial for language acquisition. For instance, graphics and dynamical visualizations may facilitate better encoding of meanings and interpretations of words (cf. dual coding theory; Clark & Paivio, 1991) or they may help learners to practice language in an authentic and playful way (e.g., the use of a massive multiplayer online role-playing game in Suh et al., 2010).

The results on the level of realism of serious games corroborate those of Vogel et al. (2006). They show that, from the perspective of learning, there is no argument to opt for photorealistic visual designs, because more basic designs such as schematic/textual and cartoonlike designs are equally or more effective. In that respect the results suggest that designers of serious games should focus more on the learning content and domain and less on visual design issues. It would be interesting to further categorize studies that we call photorealistic in photorealistic, 3-D, and virtual reality and to investigate how these levels of realism moderate learning. It would be particularly interesting when these new levels of realism are related to specific domains and types of knowledge. For example, are 3-D and virtual reality game environments more effective for learning a medical triage (the classification of victims) than a plain 2-D photorealistic game environment? For most age groups with the exception of adults, learning with serious games was more effective than conventional instruction. Vogel et al. (2006) did not find a difference between children and adults. They speculated that this was somewhat counterintuitive, given the fact that children have shorter attention spans and lower intrinsic motivation and thus may learn better than adults with computer games. Although we observed a difference, it is premature to draw a conclusion because the adults age group comprised only two comparisons.

Although serious games with a narrative are not more effective than serious games without a narrative when compared with a conventional instruction method, the difference does suggest that including a narrative is counterproductive. In this respect it seems to support the argument of Adams et al. (2012) that players will use too much of their cognitive capacity for processing the narrative information that is not directly related to the learning content. We concur with them that a story with a theme that is closely related to the learning goals may improve the effect of a narrative. Assuming that a narrative consists of a series of related events (e.g., an initiating event, exposition, complication, climax, and resolution; see Brewer & Lichtenstein, 1982), the manipulation of the order of these events may also trigger relevant cognitive

processes. The role of the manipulation of narrative events in games is still unexplored, but research on texts has shown that the introduction of surprise can be effective in terms of recall of story information and appreciation of the story (Hoeken & van Vliet, 2000). Some support for the effective use of surprising events in serious games comes from van der Spek (2011), who had learners play a narrative-based serious game in which they learned how to apply a medical procedure. During the game, specifically designed surprising events were triggered, and learners could not rely anymore on the procedure that they had learned. For example, due to a sudden failure in a power box, there was not sufficient light to perform a necessary step in the procedure. It was hypothesized that this would force players to rethink the medical procedure they had used before and to develop another solution in order to perform that step (see also Kintsch, 1980). Indeed, the unexpected events yielded a higher level of deep knowledge without a decline in the reported engagement.

We did not find statistically significant evidence for a publication bias. This is in contrast with the strong publication bias found by Sitzmann (2011) for simulation games. Possibly, the small number of unpublished pairwise comparisons (three comparisons from three studies) in our meta-analysis complicates the detection of a publication bias. We did find some evidence that the methodological rigor of the studies moderates the magnitude of the effect sizes: Designs with randomization of participants report significantly smaller effect sizes in favor of serious games than do studies with no randomization.

Motivation

Perhaps the foremost reason to use serious games is their alleged motivational appeal (Garris et al., 2002; Malone, 1981). The assumption underlying the motivational appeal of serious games is based on the high entertainment value of commercial computer games. However, the results of the meta-analysis show that serious games are not more motivating than the instructional methods used in the comparison group ($d = 0.26$, but the difference is not significant). Three plausible arguments may explain the lack of higher motivation for serious games. To start with, it is possible that serious games are not more motivating than other instructional methods. Reasoning from the self-determination approach, Ryan et al. (2006) have argued that autonomy supports intrinsic motivation. Consequently, conditions that limit the sense of control or freedom of action may undermine intrinsic motivation (Deci, Koestner, & Ryan, 1999). In serious games, the level of control is twofold: It is applicable to actions and decisions within the game but also to the instructional context, where decisions about issues such as the type of game and when to play the game have to be made. It is relevant to investigate whether variations in the level of control that serious games offer moderate intrinsic motivation. We tried to classify these variations in the studies included in this meta-analysis but found that the majority of the papers lacked sufficient information for us to do this adequately. With respect to the level control in the instructional context, an essential difference between leisure computer games and serious games is that the former are chosen by the players and played whenever and for as long as they want, whereas the type of game that is used and the playing time are generally defined by the curriculum in the case of serious games. Within the instructional context, it is possible that

the lack of control on these decisions has attenuated the motivation appeal of serious games.

The second explanation contends that the connection between game design with a focus on entertainment and instructional design with a focus on learning is not a natural one. Several dimensions that have to be resolved in order to create really engaging serious games, such as learning versus playing or freedom versus control, have been outlined (de Castell & Jenson, 2003; Wouters, van Oostendorp, Boonekamp, & van der Spek, 2011). Take the situation in which a designer uses a pop-up screen with a message that prompts the player to reflect. From an instructional design perspective such a focus may yield learning, but it is also likely that such an intervention will disturb the flow of the game and consequently undermine the entertaining nature of the game. It is plausible that the lack of motivational appeal is a reflection of the fact that the world of game design and that of instructional design are not yet integrated. If this is true, more research on factors that connect the worlds of game design and instructional design is required. Interesting in this respect is the work of Habgood and Ainsworth (2011), who found that the integration of arithmetical content with the game mechanics that make playing games entertaining was more motivating than a game version in which both components were not integrated. The third explanation stems from an examination of the methods that are commonly used for the measurement of motivation. The question can be raised whether it makes sense to measure affective states such as motivation and enjoyment with questionnaires and surveys after game play; physiological or behavioral measures such as eye tracking and skin conductance seem to be more appropriate methods, because they can be collected during game play. Also, the player's motivation during game play may be attenuated after the game has finished. In 30 of the 31 pairwise comparisons in the meta-analysis, motivation was measured with a survey or questionnaire conducted after game play. The exception (Annetta et al., 2009) used the rating of observed engagement during game play as motivation measurement and found that the game was more motivating than the instructional treatment of the comparison group who received practice and group discussion (estimated effect size $d = 0.81$).

Limitations and Directions for Future Research

Scholars have different views on what studies to include in a meta-analysis, varying from a broad sample with different study characteristics coded to a restricted sample that meets specific criteria. In this meta-analysis, we have chosen a broad focus including not only studies conducted in controlled laboratory settings but also studies that took place in a classroom setting. At the same time we have tried to further qualify the weighted mean effect size of the analysis with a number of moderators such as the methodological quality of the studies or the distribution of learning (one session vs. multiple sessions). We are aware of the fact that another view on what studies to include in the meta-analysis may lead to other conclusions regarding the effectiveness of serious games. For example, if only studies with a randomized sample and a pretest-posttest design are considered, the positive effect in favor of serious games may disappear. In addition, our selection of moderators is not exhaustive,

and other interesting features of studies (e.g., gender) may influence the effect size.

A broad range of serious games, from adventure games to puzzle games, and their application in different domains have been examined. This large variation justifies some caution when generalizing the results. The same domain can be approached from different game genres. For example, Kebritchi et al. (2010) used a sophisticated 3-D adventure game to teach mathematical skills, and Van Eck and Dempsey (2002), in the same domain, used a basic simulation game. Despite the different game genres that were used, both studies contributed to the $d = 0.17$ for mathematics. It would be interesting to investigate whether specific game genres (e.g., adventure games, simulation games) are more apt to teach specific domains (e.g., mathematics).

Our results corroborate other findings indicating that serious games are a more effective than other instruction methods (cf. Sitzmann, 2011; Vogel et al., 2006). The next step is more value-added research on specific game features that determine this effectiveness. Given the increasing number of empirical studies with serious games, we believe, a meta-analysis on serious game features can be successful. An example is the role of competition, which is regarded by some scholars as a crucial characteristic of computer games (see the introduction), but the question is whether competition is required to make effective and compelling serious games. Our review of the literature revealed some studies comparing competition and noncompetition game versions (Ke, 2008; Ke & Grabowski, 2007; Van Eck & Dempsey, 2002) that warrant such an investigation. Also, from a cognitive consequences approach, there are interesting directions for future research. For example, we found many studies investigating the effect of playing computer games on basic cognitive abilities, such as visual attention and spatial ability. We did not take these studies into account, because the “no activity” control group in these studies did not meet our inclusion criteria. With a sample of 17 comparisons we found a $d = 0.33$, indicating that computer games are effective to train basic cognitive skills. Assuming that these basic cognitive skills are associated with cognitive skills such as problem solving, it would be interesting to examine whether serious games foster these cognitive processes and whether training of these processes also yields a better performance on cognitive skills such as problem solving.

Besides the issue of the method of measurement of motivation that we addressed earlier, the definition of motivation should be examined. We applied a broad definition of motivation, which included engagement, interest, enjoyment, the ARCS (Attention Relevance Confidence Satisfaction; see Bai et al., 2012) and ATMI (Attitude Towards Mathematics Inventory; see Ke, 2008) scales, and the attitude of the player toward school or a school domain. The question can be raised whether all these definitions indeed refer to motivation or whether they represent different constructs. For example, to what extent does attitude toward school (Miller & Robertson, 2010) reflect dimensions of the construct motivation?

The conclusion that we have drawn from these results is that specific instructional or contextual features, such as supplementing with other instructional methods and working in groups, increase the effect of serious games. We have suggested that these features may have enabled learners to engage in

learning activities from which they would otherwise refrain. More research is required if these features indeed foster additional learning activities (e.g., with think-aloud protocols). And, if this is true, can we design serious games in such a way that these learning activities are also activated in stand-alone serious games or when learners play solitary games? In other words, can we design serious games in such a way that players are automatically prompted to reflect on their performance during game play?

References

References marked with an asterisk indicate studies included in the meta-analysis.

- Abdi, H. (2010). Holm's sequential Bonferroni procedure. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 574–578). Thousand Oaks, CA: Sage.
- *Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology*, 104, 235–249. doi:10.1037/a0025595
- *Annetta, L. A., Minogue, J., Holmes, S. Y., & Chen, M.-T. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education*, 53, 74–85. doi:10.1016/j.compedu.2008.12.020
- Baek, Y. K. (2008). What hinders teachers in using computer and video games in the classroom? Exploring factors inhibiting the uptake of computer and video games. *CyberPsychology & Behavior*, 11, 665–671. doi:10.1089/cpb.2008.0127
- *Bai, H., Pan, W., Hirumi, A., & Kebritchi, M. (2012). Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. *British Journal of Educational Technology*, 43, 993–1003. doi:10.1111/j.1467-8535.2011.01269.x
- Bandura, A. (1976). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.
- *Barab, S., Pettyjohn, P., Gresalfi, M., Volk, C., & Solomou, M. (2012). Game-based curriculum and transformational play: Designing to meaningfully positioning person, content, and context. *Computers & Education*, 58, 518–533. doi:10.1016/j.compedu.2011.08.001
- *Barab, S. A., Scott, B., Siyahhan, S., Goldstone, R., Ingram-Goble, A., Zuiker, S. J., & Warren, S. (2009). Transformational play as a curricular scaffold: Using videogames to support science education. *Journal of Science Education and Technology*, 18, 305–320. doi:10.1007/s10956-009-9171-5
- *Barlett, C. P., Vowels, C. L., Shanteau, J., Crow, J., & Miller, T. (2009). The effect of violent and non-violent computer games on cognitive performance. *Computers in Human Behavior*, 25, 96–102. doi:10.1016/j.chb.2008.07.008
- *Betz, J. A. (1996). Computer games: Increase learning in an interactive multidisciplinary environment. *Journal of Educational Technology Systems*, 24, 195–205. doi:10.2190/119M-BRMU-J8HC-XM6F
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.
- Brewer, W. F., & Lichtenstein, E. H. (1982). Stories are to entertain: A structural-affect theory of stories. *Journal of Pragmatics*, 6, 473–486. doi:10.1016/0378-2166(82)90021-2
- *Brom, C. M., Preuss, M., & Klement, D. (2011). Are educational computer micro-games engaging and effective for knowledge acquisition at high-schools? A quasi-experimental study. *Computers & Education*, 57, 1971–1988. doi:10.1016/j.compedu.2011.04.007

- *Cameron, B., & Dwyer, F. (2005). The effect of online gaming, cognition and feedback type in facilitating delayed achievement of different learning objectives. *Journal of Interactive Learning Research*, 16, 243–258.
- Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3, 149–210. doi:10.1007/BF01320076
- Clark, R. E., Yates, K., Early, S., & Moulton, K. (2010). An analysis of the failure of electronic media and discovery-based learning: Evidence for the performance benefits of guided training methods. In K. H. Silber & R. Foshay (Eds.), *Handbook of training and improving workplace performance: Vol. I. Instructional design and training delivery* (pp. 263–297). Somerset, NJ: Wiley.
- Cordova, D. L., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88, 715–730. doi:10.1037/0022-0663.88.4.715
- Day, E. A., Arthur, W., Jr., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology*, 86, 1022–1033. doi:10.1037/0021-9010.86.5.1022
- de Castell, S., & Jenson, J. (2003). Serious play. *Journal of Curriculum Studies*, 35, 649–665. doi:10.1080/0022027032000145552
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125, 627–668. doi:10.1037/0033-2909.125.6.627
- *Dede, C., Clarke, J., Ketelhut, D. J., Nelson, B., & Bowman, C. (2005, April). *Students' motivation and learning of science in a multi-user virtual environment*. Paper presented at the meeting of the American Educational Research Association, Montréal, Quebec, Canada.
- Dickey, M. D. (2006). Game design narrative for learning: Appropriating adventure game design narrative devices and techniques for the design of interactive learning environments. *Educational Technology Research and Development*, 54, 245–263. doi:10.1007/s11423-006-8806-y
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, England: Cambridge University Press.
- Fletcher, J. D., & Tobias, S. (2006, February). *Using computer games and simulations for instruction: A research review*. Paper presented at the Conference on New Learning Technologies, Society for Applied Learning Technology, Orlando, FL.
- Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, 33, 441–467. doi:10.1177/1046878102238607
- Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, 15, 313–331. doi:10.1016/j.learninstruc.2005.07.001
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Thousand Oaks, CA: Sage.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371–395. doi:10.1037/0033-295X.101.3.371
- *Gremmen, H., & Potters, J. (1997). Assessing the efficacy of gaming in economic education. *Journal of Economic Education*, 28, 291–303.
- Habgood, M. P. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *Journal of the Learning Sciences*, 20, 169–206. doi:10.1080/10508406.2010.508029
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hoeken, H., & van Vliet, M. (2000). Suspense, curiosity, and surprise: How discourse structure influences the affective and cognitive processing of a story. *Poetics*, 27, 277–286. doi:10.1016/S0304-422X(99)00021-2
- Inkpen, K., Booth, K. S., Klawe, M., & Uptis, R. (1995). *Playing together beats playing apart, especially for girls*. Paper presented at the meeting of Computer Support for Collaborative Learning '95, Bloomington, IN.
- *Jarvis, S., & de Freitas, S. (2009). Evaluation of an immersive learning programme to support triage training. *Proceedings of the First IEEE International Conference in Games and Virtual Worlds for Serious Applications* (pp. 117–122). doi:10.1109/VS-GAMES
- *Jong, M. S. Y., Shang, J. J., Lee, F. L., Lee, J. H. M., & Law, H. Y. (2006). Learning online: A comparative study of a situated game-based learning approach and a traditional web-based approach. In Z. Pan, R. Aylett, H. Diener, X. Jin, S. Göbel, & L. Li (Eds.), *Lecture notes in computer science: Vol. 3942. Technologies for e-learning and digital entertainment* (pp. 541–551). Berlin, Germany: Springer.
- *Ke, F. (2008). Computer games application within alternative classroom goal structures: Cognitive, metacognitive, and affective evaluation. *Education Technology Research and Development*, 56, 539–556. doi:10.1007/s11423-008-9086-5
- Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R. E. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (Vol. 1, pp. 1–32). Hershey, PA: Information Science Reference.
- *Ke, F., & Grabowski, B. (2007). Gameplaying for maths learning: Cooperative or not? *British Journal of Educational Technology*, 38, 249–259. doi:10.1111/j.1467-8535.2006.00593.x
- *Kebritchi, M., Hirumi, A., & Bai, H. (2010). The effect of modern computer games on mathematics achievement and class motivation. *Computers & Education*, 55, 427–443. doi:10.1016/j.compedu.2010.02.007
- Kintsch, W. (1980). Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics*, 9, 87–98. doi:10.1016/0304-422X(80)90013-3
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.
- Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, 29, 133–159. doi:10.1016/0749-596X(90)90069-C
- Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78, 311–328. doi:10.1037/0021-9010.78.2.311
- *Kuo, M.-J. (2007, March). How does an online game based learning environment promote students' intrinsic motivation for learning natural science and how does it affect their learning outcomes? In T.-W. Chan, A. Paiva, D. W. Shaffer, Kinshuk, & J.-C. Yang (Eds.), *The First IEEE International Workshop on Digital Game and Intelligent Toy Enhanced Learning, 2007. DIGTEL '07*, 135–142. doi:10.1109/DIGTEL.2007.28
- *Laffey, J. M., Espinosa, L., Moore, J., & Lodree, A. (2003). Supporting learning and behavior of at-risk young children: Computers in urban education. *Journal of Research on Technology in Education*, 35, 423–440.
- Leemkuil, H., & de Jong, T. (2011). Instructional support in games. In S. Tobias & D. Fletcher (Eds.), *Computer games and instruction* (pp. 353–369). Charlotte, NC: Information Age.
- Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 5, 333–369. doi:10.1207/s15516709cog0504_2
- Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press.
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, 59, 14–19. doi:10.1037/0003-066X.59.1.14
- Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 281–305). Charlotte, NC: Information Age.

- Mayer, R. E., Griffith, E., Naftaly, I., & Rothman, D. (2008). Increased interestingness of extraneous details leads to decreased learning. *Journal of Experimental Psychology: Applied*, 14, 329–339. doi:10.1037/a0013835
- *McQuiggan, S., Rowe, J., Lee, S., & Lester, J. (2008). *Story-based learning: The impact of narrative on learning experiences and outcomes*. Montreal, Quebec, Canada.
- *Miller, D. J., & Robertson, D. P. (2010). Using a games console in the primary classroom: Effects of brain training programme on computation and self-esteem. *British Journal of Educational Technology*, 41, 242–255. doi:10.1111/j.1467-8535.2008.00918.x
- *Miller, D. J., & Robertson, D. P. (2011). Educational benefits of using game consoles in a primary classroom: A randomised controlled trial. *British Journal of Educational Technology*, 42, 850–864. doi:10.1111/j.1467-8535.2010.01114.x
- Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, 97, 117–128. doi:10.1037/0022-0663.97.1.117
- *Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents. *Cognition and Instruction*, 19, 177–213. doi:10.1207/S1532690XCI1902_02
- *Nicol, M. M., & Anderson, A. (2000). Computer-assisted vs. teacher-directed teaching of numeracy in adults. *Journal of Computer Assisted Learning*, 16, 184–192. doi:10.1046/j.1365-2729.2000.00131.x
- *Okolo, C. M. (1992). The effect of computer-assisted instruction format and initial attitude on the arithmetic facts proficiency and continuing motivation of students with learning disabilities. *Exceptionality*, 3, 195–211. doi:10.1080/09362839209524815
- O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *Curriculum Journal*, 16, 455–474. doi:10.1080/09585170500384529
- *Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education*, 52, 1–12. doi:10.1016/j.compedu.2008.06.004
- *Parchman, S. W., Ellis, J. A., Christinaz, D., & Vogel, M. (2000). An evaluation of three computer-based instructional strategies in basic electricity and electronics. *Military Psychology*, 12, 73–87. doi:10.1207/S15327876MP1201_4
- Pierfy, D. A. (1977). Comparative simulation game research: Stumbling blocks and stepping. *Simulation & Games*, 8, 255–268. doi:10.1177/003755007782006
- Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.
- Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology*, 14, 154–166. doi:10.1037/a0019440
- Renkl, A., & Atkinson, R. K. (2002). Learning from examples: Fostering self-explanations in computer-based learning environments. *Interactive Learning Environments*, 10, 105–119. doi:10.1076/1052.105.7441
- *Ricci, K. E., Salas, E., & Cannon-Bowers, J. A. (1996). Do computer-based games facilitate knowledge acquisition and retention? *Military Psychology*, 8, 295–307. doi:10.1207/s15327876mp0804_3
- *Ritterfeld, U., Shen, C., Wang, H., Nocera, L., & Wong, W. L. (2009). Multimodality and interactivity: Connecting properties of serious games with educational outcomes. *CyberPsychology & Behavior*, 12, 691–697. doi:10.1089/cpb.2009.0099
- *Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., . . . Salinas, M. (2003). Beyond Nintendo: Design and assessment of educational video games for first and second grade students. *Computers & Education*, 40, 71–94. doi:10.1016/S0360-1315(02)00099-4
- Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, 118, 183–192. doi:10.1037/0033-2909.118.2.183
- Roy, M., & Chi, M. T. H. (2005). The self-explanation principle in multimedia learning. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 271–286). New York, NY: Cambridge University Press.
- Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, 30, 347–365. doi:10.1007/s11031-006-9051-8
- *Seelhammer, C., & Niegemann, H. M. (2009). Playing games to learn—Does it actually work? *Proceedings of the 17th International Conference on Computers in Education* (pp. 675–681). Hong Kong: Asia-Pacific Society for Computers in Education.
- *Segers, E., & Verhoeven, L. (2003). Effects of vocabulary training by computer in kindergarten. *Journal of Computer Assisted Learning*, 19, 557–566. doi:10.1046/j.0266-4909.2003.00058.x
- *Sindre, G., Natvig, L., & Jahre, M. (2009). Experimental validation of the learning effect for a pedagogical game on computer fundamentals. *IEEE Transactions on Education*, 52, 10–18. doi:10.1109/TE.2007.914944
- Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, 64, 489–528. doi:10.1111/j.1744-6570.2011.01190.x
- *Squire, K., Barnett, B., Grant, J. M., & Higginbotham, T. (2004). Electromagnetism supercharged! Learning physics with digital simulation games. *Proceedings of the International Conference on Learning Sciences*, 6, 513–520.
- *Suh, S., Kim, S. W., & Kim, N. J. (2010). Effectiveness of MMORPG-based instruction in elementary English education in Korea. *Journal of Computer Assisted Learning*, 26, 370–378. doi:10.1111/j.1365-2729.2010.00353.x
- Tobias, S., Fletcher, J. D., Dai, D. Y., & Wind, A. P. (2011). Review of research on computer games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 127–222). Charlotte, NC: Information Age.
- van der Meij, H., Albers, E., & Leemkuil, H. (2011). Learning from games: Does collaboration help? *British Journal of Educational Technology*, 42, 655–664. doi:10.1111/j.1467-8535.2010.01067.x
- van der Spek, E. D. (2011). *Experiments in serious game design: A cognitive approach* (Unpublished doctoral dissertation). Universiteit Utrecht, Utrecht, the Netherlands.
- *van Dijk, V. (2010). *Learning the triage procedure: Serious gaming based on guided discovery learning versus studying worked examples* (Unpublished master's thesis). Universiteit Utrecht, Utrecht, the Netherlands.
- *Van Eck, R., & Dempsey, J. (2002). The effect of competition and contextualized advisement on the transfer of mathematics skills in a computer-based instructional simulation game. *Educational Technology, Research and Development*, 50, 23–41. doi:10.1007/BF02505023
- *Virvou, M., Katsionis, G., & Manos, K. (2005). Combining software games with education: Evaluation of its educational effectiveness. *Educational Technology & Society*, 8(2), 54–65.
- Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, 34, 229–243. doi:10.2190/FLHV-K4WA-WPVQ-H0YM
- Wittrock, M. C. (1974). Learning as a generative activity. *Educational Psychologist*, 11, 87–95. doi:10.1080/00461527409529129
- Wouters, P., van Oostendorp, H., Boonekamp, R., & van der Spek, E. D. (2011). The role of game discourse analysis and curiosity in creating engaging and effective serious games by implementing a back story and foreshadowing. *Interacting With Computers*, 23, 329–336. doi:10.1016/j.intcom.2011.05.001
- Wouters, P., Paas, F., & van Merriënboer, J. J. M. (2008). How to optimize learning from animated models: A review of guidelines base on cognitive load. *Review of Educational Research*, 78, 645–675. doi:10.3102/0034654308320320

Wouters, P., van der Spek, E. D., & van Oostendorp, H. (2009). Current practices in serious game research: A review from a learning outcomes perspective. In T. M. Connolly, M. Stansfield, & L. Boyle (Eds.), *Games-based learning advancements for multisensory human computer interfaces: Techniques and effective practices* (pp. 232–250). Hershey, PA: IGI Global.

Wouters, P., van der Spek, E. D., & van Oostendorp, H. (2011). Measuring learning in serious games: A case study with structural assessment. *Educational Technology Research and Development*, 59, 741–763. doi: 10.1007/s11423-010-9183-0

*Wrzesien, M., & Raya, M. A. (2010). Learning in serious virtual worlds: Evaluation of learning effectiveness and appeal to students in the E-Junior project. *Computers & Education*, 55, 178–187. doi:10.1016/j.compedu.2010.01.003

*Yip, F. W. M., & Kwan, A. C. (2006). Online vocabulary games as a tool for teaching and learning English vocabulary. *Educational Media International*, 43, 233–249. doi:10.1080/09523980600641445

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer*, 38(9), 25–32. doi:10.1109/MC.2005.297

Appendix

Procedure to Adjust Sample Size

In the formula $N_{\text{adjusted}} = [(N_{\text{experimental group}}/a) + (N_{\text{comparison group}}/b)]/c$, a is the number of comparison groups, b is the number of experimental groups with a serious game, and c is the number of dependent variables. For example, Parchman, Ellis, Christinaz, and Vogel (2000) used two different learning outcomes ($c = 2$), three comparison groups ($a = 3$), and one experimental group ($b = 1$). The number of participants was 20 in the experimental group, 13 in the drill-and-practice comparison group, 23 in the computer-based in-

struction comparison group, and 24 in the classical instruction comparison group. This means that a total of 80 learners participated in this study. The combination of learning outcomes and comparison groups yields six pairwise comparisons. For each pairwise comparison, an adjusted n was calculated based on the number of participants in the experimental and comparison groups. As shown in Table A1, the sum of the adjusted n of all pairwise comparisons equaled the total number of participants of that study.

Table A1
Example of Adjustment of Sample Size With Two Different Learning Outcomes and Three Comparison Groups

Pairwise comparison	<i>N</i> experimental group	<i>N</i> comparison group	Dependent variable	Formula	Adjusted <i>n</i>
1	20	13	Knowledge	$([20/3] + [13/1])/2$	9.83
2	20	13	Skills	$([20/3] + [13/1])/2$	9.83
3	20	23	Knowledge	$([20/3] + [23/1])/2$	14.38
4	20	23	Skills	$([20/3] + [23/1])/2$	14.38
5	20	24	Knowledge	$([20/3] + [24/1])/2$	15.33
6	20	24	Skills	$([20/3] + [24/1])/2$	15.33
				Total <i>N</i>	80

Received November 1, 2011
Revision received October 12, 2012
Accepted November 12, 2012 ■

Reducing Verbal Redundancy in Multimedia Learning: An Undesired Desirable Difficulty?

Carole L. Yue, Elizabeth Ligon Bjork, and Robert A. Bjork
University of California, Los Angeles

Previous research on the redundancy principle in multimedia learning has shown that although exact correspondence between on-screen text and narration generally impairs learning, brief labels within an animation can improve learning. To clarify and extend the theoretical and practical implications of these results, the authors of the present research examined the extent to which varying degrees of correspondence between on-screen text and narration in a multimedia lesson affects recall and transfer. In 2 experiments, college students viewed an animated and narrated PowerPoint lesson about the life cycle of a star, with different participants experiencing different degrees of correspondence between on-screen text and narration. Consistent with the redundancy principle and the dual-channel theory of multimedia learning, both experiments demonstrated impairment for on-screen text identical to the narration. As an extension of previous research on the redundancy principle, however, both experiments also demonstrated an advantage for a small amount of discrepancy between narration and on-screen text. On the other hand, too much discrepancy resulted in learning just as poor as when the on-screen text was identical to the narration. Metacognitive judgments revealed that participants tended to prefer on-screen text identical to the narration, even though recall and transfer scores showed that on-screen text worded slightly differently than the narration was better for learning. Results indicate that despite learner preferences to the contrary, slight discrepancies between on-screen text and narration can be a desirable difficulty, suggesting an extension to the redundancy principle that is consistent with the desirable difficulties framework as well as the cognitive theory of multimedia learning.

Keywords: multimedia learning, redundancy, desirable difficulties, metacognitive judgments

With the increasing availability of technology designed and targeted for educational purposes, instructors have an ever-increasing number of options for presenting scientific explanations. Students can see a multimedia presentation depicting the life cycle of a star, for example, as part of a brief in-class demonstration, or they can view it at home via the Internet. A recurring question in the design of multimedia materials is how much on-screen text—if any—should be added to animated lessons that are accompanied by a narration. In a survey conducted by Apperson, Laws, and Scepansky (2008), college students had a strong preference for having key phrases from the lecture written on PowerPoint slides (Microsoft Corp., Redmond, WA) and a modest preference for having the full text of the lecture available on the screen. Such preferences, however, run counter to results of previous research demonstrating that learning is worse when on-screen graphics are accompanied by a narration plus on-screen

text, versus a narration alone (Jamet & Le Bohec, 2007; Kalyuga, Chandler, & Sweller, 1999; Leahy, Chandler, & Sweller, 2003; Mayer, Heiser, & Lonn, 2001; Moreno & Mayer, 2002). This impairment, called the *redundancy effect*, is often explained as resulting from the unnecessary and excessive load required of working memory to integrate identical verbal information from both visual and auditory sources (Kalyuga, Chandler, & Sweller, 2004).

The foregoing explanation of the redundancy effect derives from the cognitive theory of multimedia learning (CTML; Mayer, 2005), which incorporates dual-channel theory (Paivio, 1986) and cognitive load theory (Sweller, van Merriënboer, & Paas, 1998). According to dual-channel theory, people have an auditory channel and a visual channel available for processing information; however, cognitive load theory stipulates that each channel has a limited capacity. Thus, when either one or both of the channels are inundated with too much material, cognitive overload occurs, leading to impaired recall or transfer of the to-be-learned material. Written words that accompany an animation and are identical to spoken words present an excess of information via the visual channel. In addition, learners expend mental resources trying to coordinate the two sources of verbal information (Sweller, 2005). Supporters of cognitive load theory would argue that this situation causes extraneous cognitive load (i.e., a characteristic of the learning environment that requires mental resources but does not contribute to learning). For these reasons, CTML posits that identical verbal information presented in two modalities creates cognitive overload in the learner and impairs learning.

This article was published Online First March 18, 2013.

Carole L. Yue, Elizabeth Ligon Bjork, and Robert A. Bjork, UCLA Bjork Learning & Forgetting Lab, Department of Psychology, University of California, Los Angeles.

This research was supported by Grant 29192G from the James S. McDonnell Foundation. We thank Kou Murayama and other members of CogFog for helpful comments regarding this research.

Correspondence concerning this article should be addressed to Carole L. Yue, UCLA Bjork Learning & Forgetting Lab, Department of Psychology, 1285 Franz Hall, Box 951563, University of California, Los Angeles, CA 90095-1563. E-mail: caroleleigh@ucla.edu

Although presenting on-screen text that is completely identical to narration appears to impair learning, a recent meta-analysis by Adesope and Nesbit (2012) of redundancy in multimedia environments found that computer-paced presentations with a low degree of correspondence (i.e., on-screen text that contains abridgments or a few key terms from the narration) generally result in better learning than presentations with a high degree of correspondence (i.e., on-screen text that is identical to the narration). One example of a low degree of correspondence that proved to be helpful comes from a study by Mayer and Johnson (2008) in which a multimedia lesson about lightning formation either included, or did not include, labels consisting of two or three words from the narration that appeared next to relevant portions of the image. Participants who viewed lessons with these labels performed better on a recall test than those who did not. This finding suggests that under certain circumstances, a minimal amount of on-screen text can induce germane load (i.e., cognitive effort that promotes learning) rather than extraneous (i.e., cognitive effort that is unrelated to the learning endeavor; Mayer & Johnson, 2008; Sweller et al., 1998) load.

There are at least two reasons why a low degree of correspondence between aural and visual text could lead to improved learning: First, highlighting only key words and phrases may help focus the reader's attention to the critical information contained in the narration and depicted on the screen (Adesope & Nesbit, 2012; Mayer & Johnson, 2008). As Mayer and Johnson hypothesized and found, highlighting those key phrases led to improved recall, apparently because the labels emphasized the key information. Second, it seems possible that a slight discrepancy between the two sources of verbal information could lead to deeper processing overall. Possibly, for example, when learners realize that what they are reading and what they are hearing are not the same at the surface level, they may be motivated to compare the two sources at a deeper level to ensure that they match conceptually (Kintsch & van Dijk, 1978; Schnotz & Bannert, 2003). Although making such a comparison would require more effortful mental processing, it would also require the learner to engage more actively with the lesson content, making deeper understanding and long-term retention more likely (Mayer, 2005). As long as accomplishing this comparison did not exceed working memory capacity, the presentation of information in this manner could function as an example of a desirable difficulty—a condition of learning that creates difficulties for the learner during acquisition or study but then actually promotes long-term retention and transfer (Bjork, 1994). In terms of cognitive load theory, this type of processing would induce germane cognitive load. If the second possibility is true, there may be some level of redundancy that promotes learning without causing cognitive overload.

A logical middle ground between two-word labels and full identical text is a shorter sentence that summarizes the narration. In Experiment 2 of Mayer et al. (2001), one group of participants experienced such a condition: They viewed an animated, narrated lesson about lightning formation and saw a short sentence on the screen that summarized the narration for each frame. Another group saw a full sentence on the screen that was identical to the narration, and a control group saw no on-screen text. On the recall test, the control group significantly outperformed the identical text group, but the performance of the summary text group did not differ significantly from either of the other groups. On the transfer

test, the control group outperformed both groups that were given either type of on-screen text, and while the summary text group performed numerically better than the identical text group, performance of the two groups was not significantly different.

The question remains, then, whether there is a way for on-screen text to promote deeper processing and lead to increased learning from a multimedia lesson. Even though the on-screen summary did not significantly help participants in Mayer et al.'s (2001) Experiment 2, it may be that a different type of abridgment, such as one that retains more of the information necessary for transfer, could be helpful. More specifically, our thinking was that if the non-matching abridgments were constructed such that they would require the learner to engage in deeper processing in order to compare the on-screen text and narration at a conceptual level, then such a presentation might be beneficial. In short, this type of verbal redundancy might function as a desirable difficulty, improving both retention and transfer of the to-be-learned information.

We addressed this question in the present Experiment 1 by conducting a conceptual replication of Mayer et al.'s (2001) Experiment 2 using nonmatching summaries of the narration designed to force learners to engage in deeper processing by conceptually comparing the on-screen text and the narration. Additionally, the lesson we employed was slightly longer than the one used by Mayer et al., and we also included a condition meant to simulate a learning situation that is becoming increasingly common in universities: an audio podcast. Our goal was to see if this level of verbal redundancy might act as a desirable difficulty, rather than create extraneous cognitive load.

Experiment 1

In Experiment 1, students viewed a multimedia presentation about the life cycle of a star; this presentation either had (a) no added on-screen text, (b) abridged on-screen text, or (c) on-screen text identical to the narration. Additionally, a fourth group (the podcast group) heard only the narration. Based on CTML, we anticipated that the podcast group would perform most poorly, as they only received information in one modality and did not have the benefit of animation. Additionally, we expected the abridged-text group to surmount the desirable difficulty of nonmatching text by engaging in deeper processing of the information and, thus, to develop a greater conceptual understanding of the lesson compared with the other groups.

Method

Participants. A total of 107 college students (77 women, 28 men; average age: 20.5 years) at a large public university participated for credit in a psychology or linguistics course. Two participants were eliminated due to their having prior knowledge of the lesson content (as measured by a score of at least 7 out of 10 on the pretest or at least 4 out of 5 on a posttest self-report scale), which left 105 participants in the final analysis: 27 in the control condition, 25 in the identical-text condition, 28 in the abridged-text condition, and 25 in the podcast condition. A total of 84 participants reported their primary language as English, and their average self-rated English proficiency was 4.44 out of 5, which did not vary across conditions, $F(3, 101) < 1$.

Materials and design. Each lesson was presented on a 21.5-in. iMac computer (Apple Inc., Cupertino, CA) and consisted

of a 253-s, system-paced PowerPoint slide show accompanied by a narration (501 words) about the life cycle of a star. The lesson contained 14 total idea units, which are listed in Appendix A and were presented across a total of nine slides, all of which are illustrated in Figure 1. The control condition consisted of an animation accompanied by the narration. The identical-full-text condition included the same narration and animation, with identical text appearing at the bottom of the screen. The abridged-text condition presented the same narration and animation, but the text at the bottom of the screen was a shortened version of the narration. The abridged text shown with each slide preserved the basic information necessary to understand the process of a star's life cycle illustrated in that slide but with any nonessential information removed. To induce conceptual comparisons instead of word-by-word comparisons, we used similar—but not exact—phrasings that captured the main idea of the narration segment for each slide. For example, the narration in the first segment was “Stars are born out of nebulae, which are clouds in space made up of dust and gas.” In contrast, the corresponding abridged version of this information was “Stars begin in nebulae, which are clouds of dust and gas.” As illustrated in this example, the abridged text was always

slightly shorter (e.g., 11 words compared with 17) but nonetheless contained the same key information. In addition, as also illustrated in this example, the slight change in wording from the narration to the abridged text (e.g., from “Stars are born in nebulae” to “Stars begin in nebulae”) always required the learner to compare their meanings (e.g., of “born in” to “begin in”) to realize that the two sources of verbal information did, in fact, communicate the same essential information.

In the abridged-text and identical-full-text conditions, the on-screen text appeared at the bottom of the screen on each slide of the animation (see Figure 1 for sample screenshots without on-screen text). The captions were visible for the full duration of that slide; they disappeared when the narrator finished that segment and moved on to the next one. The podcast condition contained only the narration accompanied by a black screen. In conditions with animation present, the images and motion on the screen were presented simultaneously with the relevant portion of the narration.

Participants took a 10-question pretest before the lesson to assess their prior knowledge of stars. The pretest included questions such as “What is the first stage in the life cycle of a star?” and “What causes high luminosity of a star?” The posttest following

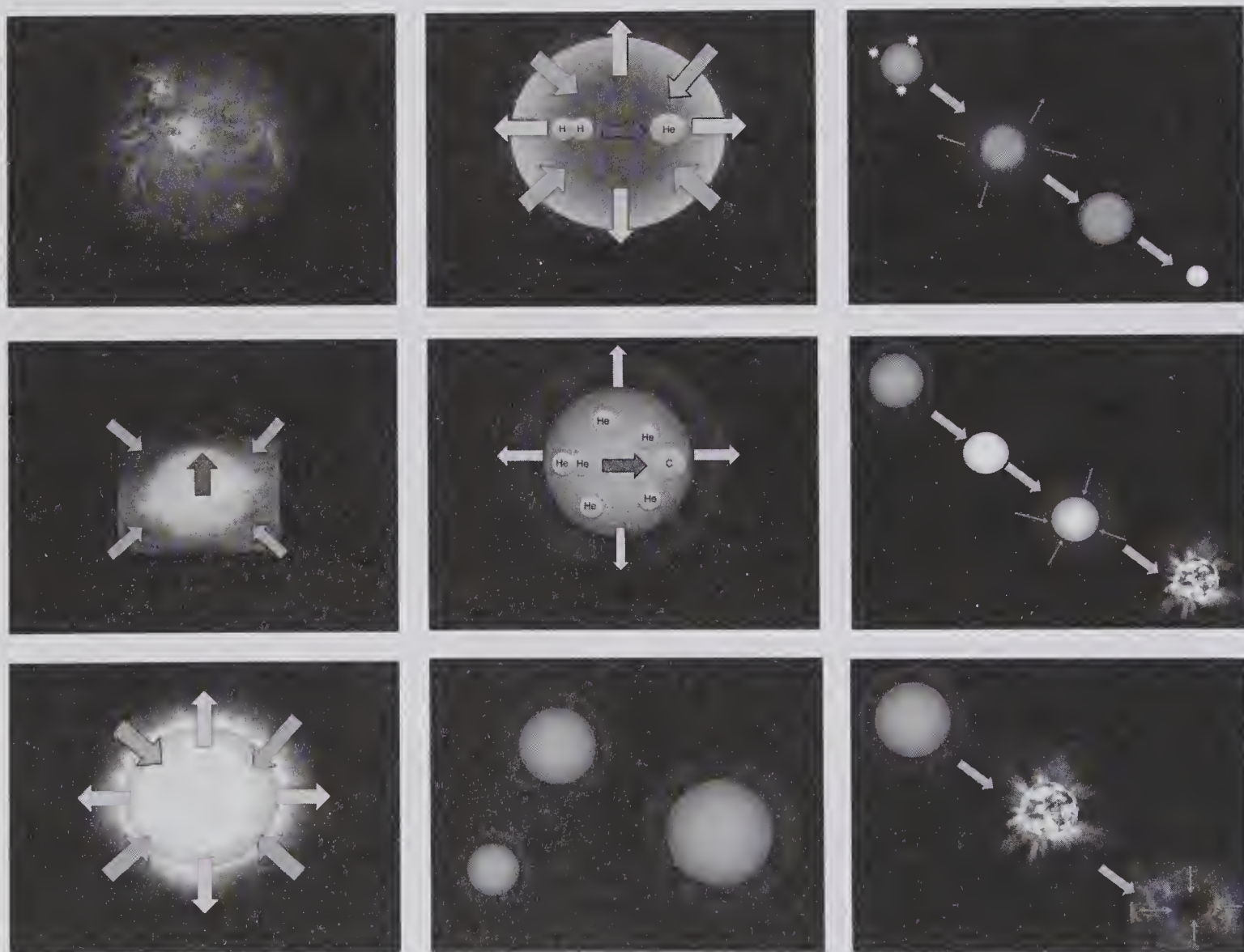


Figure 1. Screenshots of each slide of the animation for all conditions. In conditions with text, the words appeared as captions at the bottom of the screen.

the lesson was administered via a computer-based form on which participants typed in answers to a free recall question ("Please describe the life cycle of a star in as much detail as you can remember") and four transfer questions (see Appendix B for questions and possible answers). The answers to the transfer questions were not explicitly stated in the lesson, but they could be inferred from the information presented. For example, the presentation stated that a star was kept in equilibrium by a balance of gas pressure pushing outward and gravity pulling inward. In a separate segment, the presentation also stated that stars expand during the red giant phase. Therefore, if participants combined those two pieces of information, they could infer that gas pressure exceeds gravity in the red giant phase (a correct answer to the third transfer question). A stopwatch was used to record the time participants spent on the recall and transfer portions of the test.

The posttest also included two metacognitive questions: "If you had an option, what type of presentation would you prefer to see?" and "Which type of presentation do you think would result in the best learning?" Because each participant only experienced one condition, the following answer choices were provided for each of these questions so that participants could judge among them: (a) images and narration only; (b) images, narration, and on-screen text identical to the narration; (c) images, narration, and on-screen text summarizing the narration; and (d) podcast style—narration only, without images. A demographic questionnaire also was used to collect information on gender, age, and self-rated English proficiency. Finally, in case the pretest was too hard or some participants were avoiding guessing, we also asked participants how much of the information from the lesson, on a scale of 1 (*none of it*) to 5 (*all of it*), they knew prior to the experiment.

Procedure. Participants were randomly assigned to groups and tested individually in a quiet room. They were given the pretest and then asked to put headphones on before the experimenter began the lesson on the computer. After the lesson, participants took an immediate free-recall test, followed by the four transfer questions. Participants first typed in their answer to the free-recall question and then clicked a *Continue* button to move on to the transfer questions. All four of the transfer questions were presented on the screen at once. Participants had as much time as they needed to complete the test, and their response times for recall and transfer were recorded separately. After the transfer test, participants answered the metacognitive and demographic questions on the computer.

Results and Discussion

All means are reported as the proportion correct out of 14 (recall) or 4 (transfer), and all effect sizes are reported in terms of partial eta squared for analyses of variance (ANOVAs) and Cohen's d for t tests. Two raters independently scored participants' recall and transfer responses. Although scoring was dichotomous (i.e., correct or incorrect), the scoring procedure was relatively lenient: Regardless of wording, we accepted any response that indicated the participant recalled the main point of that idea unit. For example, for the idea unit, "Stars are born out of nebulae, which are clouds in space made up of dust and gas," we accepted responses such as "A star first forms from dust and gas particles." We were similarly lenient on transfer answers: Any logical response received a score of 1. We accepted responses such as "Not having a strong enough gravitational pull" or "Ability to fuse heavier elements" for the transfer question, "What could prevent a medium- or high-mass star from going supernova?" Interrater reliability was .92, and discrepancies were discussed and resolved.

Prior-knowledge assessments. Participants with too much prior knowledge, as indexed by predetermined criteria (scoring 7 or higher on the pretest or 4 or higher on the posttest knowledge self-report) were eliminated. The pretest scores for the remaining participants were low, ranging from 0–5, with a median score of 0 and an average score of 0.75 out of 10. One-way ANOVAs indicated no significant differences in pretest scores across conditions, $F(3, 101) = 2.32$, mean square error (MSE) = 2.62, $p > .05$, or in responses to the knowledge self-report question ($M = 1.26$), $F(3, 101) = 1.38$, $MSE = 0.33$, $p > .05$.

Recall. The recall scores across the four conditions are shown in the left column of Table 1. A between-subjects ANOVA revealed a significant main effect of condition on recall performance, $F(3, 101) = 14.02$, $MSE = 0.37$, $p < .001$, $\eta_p^2 = .29$. Consistent with CTML, post hoc comparisons using Fisher's least significant difference (LSD) test indicated that recall was significantly better in the three conditions with animation—control ($p < .001$, $d = 1.26$), identical-full-text ($p = .004$, $d = 1.00$), and abridged-text ($p < .001$, $d = 1.90$)—than in the podcast condition. In line with previous research on the redundancy principle, participants in the control condition recalled marginally more than participants in the identical-full-text condition ($p = .06$, $d = 0.48$). Additionally, in support of a desirable difficulties explanation, participants in the abridged-text group recalled significantly more than those in the identical-full-text group ($p = .002$, $d = 0.95$). Although partici-

Table 1
Average Proportion Correct Recall and Transfer in Final Test Performance in Experiment 1

Experimental condition	Recall		Transfer	
	M (SD)	Effect size (d)	M (SD)	Effect size (d)
Control	.33 (.21)	—	.34 (.26)	—
Identical full text	.25 (.14)	0.48	.36 (.30)	0.07
Abridged text	.39 (.15)	0.33 [†]	.50 (.27)	0.6*
Podcast	.11 (.13)	1.9*	.27 (.24)	0.28

Note. Recall was scored out of 14, and transfer was scored out of 4. Effect sizes are reported as Cohen's d compared with the control group. An * indicates a significant difference from the control group at $p < .05$.

[†] $p = .06$.

pants in the abridged-text group also scored numerically better on the recall test than participants in the control group, that difference was not significant ($p = .22$). Participants took an average of 234.7 s on the free recall portion of the test, and time did not vary across condition, $F(3, 101) = 2.11$, $MSE = 37687.45$, $p > .05$.

Transfer. As indicated in the third column of Table 1, a significant main effect of condition on transfer performance was also obtained, $F(3, 101) = 3.45$, $MSE = 0.25$, $p = .02$, $\eta_p^2 = .09$. Once again, consistent with the desirable-difficulties idea, Fisher’s LSD post hoc comparisons revealed that participants in the abridged-text group performed significantly better on the transfer questions than did participants in the control group ($p = .03$, $d = 0.60$) and the podcast group ($p = .002$, $d = 0.90$) and marginally better than did participants in the identical-full-text group ($p = .06$, $d = 0.50$). Thus, all of these comparisons are in the direction hypothesized and support a desirable-difficulties explanation. Nonetheless, the marginal nature of the third comparison indicates a need for caution. Participants took an average of 171.8 s for all four transfer questions, and time did not vary by condition, $F(3, 101) = 1.61$, $MSE = 14328.35$, $p > .05$, $\eta_p^2 = .05$.

Metacognitive judgments. As shown in Table 2, participants never stated that they would prefer the podcast condition, and they rarely stated that they would prefer viewing an animation and narration without on-screen text. In addition, participants who experienced either of the on-screen text conditions (identical-full-text or abridged-text) tended to indicate a preference for the identical-full-text condition and to think it would be the best for learning. Those in either the control or the podcast groups, however, thought that the abridged-text condition would be the most preferred and the best for learning. That is, whereas participants receiving any amount of on-screen text tended to believe that receiving on-screen text identical to the narration would be best for learning, participants not receiving any type of on-screen text tended to judge that receiving abridged text would be best for learning. A 2 (text condition: present vs. absent) \times 2 (preferred condition: identical-full-text vs. abridged-text) chi square test of independence revealed that this pattern was indeed significant, $\chi^2(1, N = 96) = 20.15$, $p < .001$, meaning that when any amount of on-screen text had been present during learning, students said they would prefer to see identical on-screen text, but when on-screen text had been absent, participants thought they would prefer abridged on-screen text. The same significant pattern of results occurred for the judgment of which condition would be best for learning, $\chi^2(1, N = 98) = 10.63$, $p = .001$.

The pattern of results in Table 2 seems consistent with prior research involving metacognitive judgments about desirable diffi-

culties (e.g., Karpicke, 2009; Kornell & Bjork, 2008): namely, that students who experience greater, but beneficial, challenges during the learning phase (e.g., having to reconcile discrepant forms of verbal information) tend not to appreciate the benefits of such challenges during acquisition, and students who feel they have experienced a fluent condition (e.g., matching text and narration in this situation) tend not to think that a less fluent one would be beneficial.

Experiment 2

In Experiment 1, we partially replicated the redundancy effect: Seeing on-screen text identical to the narration led to worse recall and transfer than did having no on-screen text. Additionally, and consistent with the desirable-difficulties hypothesis, we also found that the lesson with on-screen abridged text led to better transfer performance than the control group. However, participants who received abridged text thought that having identical-full-text captions would be best for learning, whereas students who did not receive any on-screen text thought receiving abridged text would be best. This difference in preferences could be due to participants’ differing impressions of how much text the “abridged” version would actually contain, or perhaps some participants may not have realized that the abridged text would be on the screen at the same time as the animation. We attempted to clarify the metacognitive questions in Experiment 2 to account for such possibilities.

Several other considerations motivated the design of Experiment 2. One was that our abridged-text condition acted more like Mayer and Johnson’s (2008) nonredundant condition than Mayer et al.’s (2001) summary condition. Mayer and Johnson (2008) found that two- or three-word labels next to the relevant portion of an animation helped learning, possibly by directing the learner’s attention to the most important part of the lesson. Perhaps our abridged sentences served a similar purpose, given the increased length of our lesson, which was almost twice as long (253 s vs. 140 s) as the lesson employed by Mayer et al. (2001). Possibly, as lessons increase in length, some amount of on-screen text can help focus the learner and highlight important information from the narration.

A related consideration is that a long presentation of spoken information may induce excessive working memory load and thus impair learning, a phenomenon recently described as the *transient information effect* (Leahy & Sweller, 2011; Sweller, Ayres, & Kalyuga, 2011). Written text provides learners with a source of information that is more substantial than transient auditory information, an important consideration when lesson content is com-

Table 2
Number of Participants Preferring Each Condition (“Best for Learning” Judgment in Parentheses) in Experiment 1

Experimental condition	Condition selected			
	Control	Identical full text	Abridged text	Podcast
Control	0 (0)	8 (7)	19 (20)	0 (0)
Identical full text	3 (2)	16 (13)	6 (10)	0 (0)
Abridged text	3 (1)	18 (18)	7 (9)	0 (0)
Podcast	3 (4)	5 (7)	17 (14)	0 (0)
Total	9 (7)	47 (45)	49 (53)	0 (0)

plex. The transient information effect might help account for why our abridged-text condition was better than our control condition, but it would also predict a benefit for our identical-full-text condition, which we did not observe.

Another consideration that motivated the design of Experiment 2 was that the superior performance in the abridged-text condition versus the identical-full-text condition might have occurred because our abridged-text guided learners' cognitive processing in such a way that they simply had less to remember. Although a possible explanation of the pattern we observed, such an interpretation seems unlikely. The abridged text actually contained only 12 of the 14 total idea units presented in the narration and the identical text, meaning that any abridged-text participant relying only on the on-screen text would have been at a disadvantage, as recall scores were based on idea units recalled out of the 14 in the total narration.

In addition, the high performance on transfer questions for participants in the abridged-text condition indicates that they were paying attention to the transitions and linking phrases present in the narration. If lower cognitive load had been responsible for the better performance of the abridged-text group, then the control group, which experienced even less load than the abridged-text group, should have shown enhanced performance over the abridged-text group.

An explanation that we favor for our results—especially the superiority of the abridged-text condition—is that the mismatch between the on-screen text and the narration in the abridged-text condition functioned as a desirable difficulty for learners. Because the two sources of verbal information did not match up exactly, participants had to pay attention to both sources to reconcile the two and ensure that they contained the same information (Schnotz & Bannert, 2003). If this explanation is valid, then it suggests an extension to the redundancy principle: namely, that nonmatching on-screen text, rather than necessarily creating a redundancy effect, can induce cognitive processes that enhance learning. In the present Experiment 2, we tested this possibility by using different levels of correspondence between narration and on-screen text. To do so, we employed three new types of presentation groups in addition to the previously used identical-full-text and abridged-text presentation groups of Experiment 1.

Two of the new groups, which we refer to as a *near-change group* and a *far-change group*, were presented with full-length, but nonmatching, on-screen text. The on-screen text in both groups contained the same number of words as the narration, but the near-change group was shown text that was worded only slightly differently, while the far-change group was shown text worded very differently. If as predicted by the desirable-difficulties hypothesis nonmatching verbal information can induce deeper, germane processing of the information, then learners should benefit from on-screen text that—although containing the same number of words as the narration—uses slightly different wording to convey the same concepts. It is also possible, however, that too little correspondence between the on-screen text and narration could result in cognitive overload. Thus, these two groups were added to assess both the desirable-difficulties hypothesis and the existence of boundary conditions on the level of redundancy that might be beneficial for learning.

We considered that the near-change condition would present a similar level of desirable difficulty to that presented by the abridged-text condition of Experiment 1, and thus performance of participants in these two groups might be similar to each other and better than that

of participants in the identical-full-text group. In contrast, because their presentation had a much greater mismatch between the on-screen text and the narration, the participants in the far-change condition might require too much additional processing to compare the content of the two sources of information successfully in the limited time available, thus leading to impaired recall and transfer scores for participants in that group. In terms of the CTML, the effort required to process and reconcile the verbal information would overload working memory and cause extraneous load.

Finally, to assess further the possibility that the abridged-text advantage observed in Experiment 1 arose simply because less information was presented on the screen and thus the overall cognitive load was decreased for participants in that condition, we included an identical-abridged condition in Experiment 2 in which the narration and on-screen text were both abridged but were identical to each other. This condition gave participants less to learn than those with the full narration and full on-screen text, but still presented them with the challenge of on-screen text that was identical to the narration. If the advantage for the abridged-text condition in Experiment 1 did stem from learners having less information to process, then participants in the identical-abridged condition should perform just as well as participants in the abridged-text condition and better than the participants in the identical-full-text condition. In addition, the near-change and far-change groups should perform equally as well as participants in the identical-full-text condition and worse than those in either condition with abridgments.

Method

Participants. The participants were 159 college students (93 women, 44 men; average age: 22.1 years) at a large public university who participated for course credit in psychology or linguistics courses. Nineteen participants were eliminated due to prior knowledge of the lesson content (again, as measured by a score of at least 7 out of 10 on the pretest or at least 4 out of 5 on a posttest self-report scale), and three were eliminated due to computer error. Eliminating those participants left a total of 137 participants, 29 in the identical-full-text condition and 27 students in each of the other conditions. Participants' average self-rated English proficiency was a 4.7 out of 5, which did not vary significantly across conditions, $F(4, 132) = 1.79, p > .05$.

Materials, design, and procedure. Although the to-be-learned lesson and the presentation materials for the two conditions replicating the identical-full-text and abridged-text conditions of Experiment 1 remained the same, some modifications of materials and metacognitive questions were made to accommodate the three new conditions included in Experiment 2: near-change, far-change, and identical-abridged. Although the animation and narration remained the same for these conditions, we reworded the captions using synonyms in place of certain words in the near-change condition, and we reworded the captions so that the sentence structure differed from the narration in the far-change condition. In both conditions, however, each caption contained the same number of words and the same idea units as the narration. For example, if the narration stated, "Stars are born out of nebulae, which are clouds in space made up of dust and gas," then the near-change caption read, "Stars are created from nebulae, which are clouds made up of dust and gas in outer space." The corre-

sponding far-change caption was, "In space, clouds of dust and gas are called nebulae and are the birthplace of all stars."

In the identical-abridged condition, both the narration and the captions matched the abridged text from Experiment 1. The animation contained the same frames, but it was sped up so that the animation sequence would remain temporally contiguous with the shortened narration; thus, this lesson took less time than the others (90 s vs. 253 s).

The metacognitive questions assessing learners' preferences and judgments were revised to reflect the additional conditions of Experiment 2. The two questions asked were "If you had an option, what type of on-screen text would you prefer to accompany a presentation with images and narration?" and "What type of on-screen text do you think would result in the best learning if it were to accompany a presentation with images and narration?" As in Experiment 1, because each participant only experienced one condition, the following answer choices were provided so that participants could judge among the different conditions: (a) no on-screen text—images and narration only, (b) simultaneous on-screen text summarizing the narration, (c) simultaneous on-screen text identical to the narration, and (d) simultaneous on-screen text with the same amount of information in the narration, but worded differently.

Results

All means are reported as the proportion correct out of 14 (for recall scores) or 4 (for transfer scores), with the exception of the recall scores for the identical-abridged condition. These latter scores reflect the proportion correct out of 12, as there were only 12 idea units presented in that lesson.¹ Interrater reliability was .94, and any discrepancies between raters were discussed and resolved.

Prior knowledge assessments. After eliminating participants with a pretest score of 7 or higher or a self-report of 4 or higher, the pretest scores ranged from 0 to 4, with a median score of 0 and an average score of 0.50 out of 10. Again, there were no differences across conditions on the pretest, $F(4, 132) = 2.15$, $MSE = 1.99$, $p > .05$. The average self-reported knowledge score (1.4 out of 5) was also similar to that of Experiment 1; however, unlike Experiment 1, there were significant differences between conditions, $F(4, 132) = 3.48$, $MSE = 1.56$, $p = .01$, $\eta_p^2 = .10$. Post hoc comparisons indicated that participants in the full-identical-text group rated their knowledge ($M = 1.7$, $SE = 0.12$) as marginally higher than participants in the identical-abridged-text group ($M = 1.22$, $SE = 0.13$, $p = .06$) and significantly higher than those in the abridged-text group ($M = 1.15$, $SE = 0.13$, $p = .02$). To control for these differences, we used participants' self-rated prior knowledge as a covariate in our analyses of recall and transfer scores.

Recall. In Table 3 are shown the average recall and transfer performance scores across conditions. An analysis of covariance (ANCOVA; between-subjects factor: condition; covariate: prior knowledge) indicated that type of on-screen text affected recall, $F(4, 131) = 3.50$, $MSE = 0.124$, $p = .01$, $\eta_p^2 = .10$. Post hoc Fisher's LSD comparisons indicated that as in Experiment 1, the abridged-text condition led to significantly better recall than did the identical-full-text condition ($p = .008$, $d = 0.70$). Importantly, the near-change condition also led to marginally better recall than the identical-full-text condition ($p = .06$) with a moderate effect size ($d = 0.5$). The marginal nature of this latter result, however, indicates that this advantage for the near-change condi-

Table 3

Average Adjusted Proportion Correct Recall and Transfer in Experiment 2

Experimental condition	Final test performance	
	Recall (<i>SD</i>)	Transfer (<i>SD</i>)
Identical full text	.31 (.19)	.39 (.29)
Far change	.28 (.19)	.30 (.29)
Near change	.40 (.19)	.49 (.29)
Abridged text	.45 (.19)	.53 (.29)
Identical abridged	.37 (.19)	.31 (.29)

Note. Recall was scored out of 14 except for the identical abridged group, which was scored out of 12. All transfer scores were out of 4.

tion—although in the hypothesized direction—should be interpreted cautiously. There was no significant difference in recall between the two groups with matching on-screen text and narration (i.e., the identical-full-text and the identical-abridged-text groups, $p = .24$).

Additional post hoc comparisons confirmed that participants in the abridged-text and the near-change conditions recalled significantly more idea units than participants in the far-change condition ($p = .002$, $d = 0.8$, and $p = .02$, $d = 0.6$, respectively). Although the abridged-text and near-change groups also scored numerically better than participants in the identical-abridged-text group, these apparent advantages were not significant ($p = .14$ and $p = .32$, respectively). In addition, the far-change group did not differ from the identical-full-text group on recall ($p = .4$).

Participants took an average of 202.37 s on the free recall portion of the test and the time taken varied significantly across conditions, $F(4, 132) = 2.93$, $MSE = 60762.18$, $p = .02$, $\eta_p^2 = .08$, with participants in the abridged-text condition taking the most time ($M = 269.07$ s, $SE = 27.71$) and participants in the identical-abridged condition taking the least time ($M = 140.41$ s, $SE = 27.71$).

Transfer. The effect of condition on which transfer questions were answered correctly was not significant and, thus, the results reported in this section reflect average performance across all transfer questions.

An ANCOVA (between-subjects factor: condition; covariate: prior knowledge) revealed that the type of on-screen text did affect transfer performance, $F(4, 132) = 3.37$, $MSE = 0.28$, $p = .01$, $\eta_p^2 = .09$. Post hoc comparisons using Fisher's LSD test revealed a marginal benefit of abridged text over identical full text for participants' performance on transfer questions ($p = .08$, $d = 0.32$). Although there was a numerical advantage for the near-change condition over the identical-full-text condition, that difference did not approach significance ($p = .18$). Additionally, there were no significant differences in performance between participants in the identical-full-text group and those in the far-change group ($p = .25$) or those in the identical-abridged-text group ($p = .37$).

Participants in the abridged-text and near-change conditions scored significantly better on the transfer test than did participants in the far-change condition ($p = .004$, $d = 0.8$, and $p = .02$, $d = 0.7$, respectively). In addition, participants in both the abridged-text and near-change conditions also performed better on the

¹ When all conditions were scored out of the 12 idea units in the identical-abridged lesson, the statistical patterns remained the same.

transfer test than did participants in the identical-abridged-text condition ($p = .01$, $d = 0.7$, and $p = .03$, $d = 0.66$, respectively).

Participants took an average of 155.77 s to answer all of the transfer questions, and the time did vary across group, $F(4, 132) = 2.94$, $MSE = 20867.46$, $p = .02$, $\eta_p^2 = .08$. Just as with the recall test, participants in the abridged-text condition took the most time on the transfer test ($M = 203.89$ s, $SE = 16.21$), and participants in the far-change condition took the least amount of time ($M = 135.63$ s, $SE = 16.21$).

Metacognitive judgments. As shown in Table 4, there were few discrepancies between the condition participants preferred and the condition they judged best for learning. More than 90% of participants said they would prefer a narrated animation with some sort of on-screen text, and about half responded that they would prefer identical text. Consistent with previous research indicating that students tend not to realize the benefits of desirable difficulties during study, only about one fourth of the participants stated that they would prefer to see abridged captions, and five participants said they would prefer nonidentical full-length captions. In contrast to Experiment 1, a 5 (condition: identical-full-text vs. abridged-text vs. near-change vs. far-change vs. identical-abridged) $\times 2$ (preferred condition: abridged text vs. identical text) chi-square test of independence indicated that condition did not influence learners' preferences, $\chi^2(4, N = 119) = 5.11$, $p = .28$, or judgments of which condition would be best for learning, $\chi^2(4, N = 119) = 2.08$, $p = .72$.

Discussion

In Experiment 2, we again found benefits for abridged text over identical-full text (significantly so for recall performance and marginally so for transfer performance). We also found that the near-change condition, in which the on-screen captions were slightly different from the narration, resulted in marginally better recall than the identical-full-text condition. When the on-screen captions were too different from the narration, however, as in the far-change condition, both recall performance and transfer performance were as low as in the identical-full-text condition.

Overall, these findings seem consistent with predictions that follow from the desirable-difficulties hypothesis—namely, that the abridged-text and near-change conditions should result in the best performance, while the identical-full-text and identical-abridged conditions should result in the poorest performance. There is support, too, for the cognitive load theory, which predicts that the far-change condition

should also result in poor performance owing to the high load imposed by very different narration and on-screen text.

It is important to note that there was no advantage for simply having a shorter lesson with less information; in fact, the identical-abridged-text condition resulted in no better recall and a worse average transfer score than did the abridged on-screen text with the full narration. It is possible that because the identical-abridged condition was much shorter than the other conditions and the animations had to be sped up, participants in that condition did not have sufficient time to engage in deeper processing. On the other hand, it could also be argued that the participants in the identical-abridged condition had less of a delay to the test than did participants in the other conditions—not to mention fewer words to encode. Hence, a more likely reason for these results is that, rather than focusing only on the information presented in the abridged text, learners performed some extra processing when the on-screen text and narration did not match. The similar scores in the abridged-text and the near-change group (i.e., both groups with slight differences between on-screen text and narration) support this explanation. The poor performance in the far-change group, however, suggests that there is a limit to the level of discrepancy that is beneficial: When the on-screen text differs too much from the narration, the amount of cognitive processing required to reconcile the two sources causes cognitive overload, impairing—rather than facilitating—learning.

Furthermore, participants were unaware of the benefits of non-matching verbal information. Across all conditions, participants tended to judge that they would prefer on-screen text identical to the narration. These results are consistent with the metacognitive judgments obtained in Experiment 1: When participants experienced any sort of on-screen text, they judged that having identical text would be the most optimal learning condition.

General Discussion

An overarching goal of the present research was to determine whether low correspondence between narration and on-screen text in a multimedia lesson might function as a desirable difficulty and, thus, facilitate learning. Consistent with previous research and CTML, we found that presenting on-screen text identical to a narration resulted in worse recall and transfer than when no on-screen text was presented (Kalyuga et al., 1999, 2004; Mayer et al., 2001). We also found, however, consistent with the desirable-difficulties hypothesis, that participants generally performed better on recall and transfer tests when on-screen text varied slightly

Table 4
Number of Participants Preferring Each Condition ("Best for Learning" Judgment in Parentheses) in Experiment 2

Experimental condition	Condition selected			
	Images/no text	Identical text	Abridged text	Nonidentical full text
Identical full text	2 (1)	16 (15)	10 (11)	1 (2)
Far change	3 (3)	16 (16)	5 (7)	3 (1)
Near change	3 (2)	20 (18)	4 (6)	0 (1)
Abridged text	3 (4)	14 (13)	10 (8)	0 (2)
Identical abridged text	2 (2)	15 (16)	9 (7)	1 (2)
Total	13 (12)	81 (78)	38 (39)	5 (8)

from the narration than when text was completely absent (Experiment 1) or identical to the narration (Experiments 1 and 2). Furthermore, the general benefit for low correspondence remained when the on-screen text highlighted only key phrases from the narration and when it was the same length as the narration.

Theoretical Implications

According to the dual-channel theory and previous research on the redundancy principle in multimedia learning, words presented simultaneously in visual and auditory formats result in poorer learning when they accompany an animation (Kalyuga et al., 1999, 2004; Mayer et al., 2001; Moreno & Mayer, 2002). When narration is present, on-screen text can cause extraneous processing for two reasons: First, visually presented text can overload the visual channel by presenting too much information for visual working memory to process simultaneously (Mayer et al., 2001; Moreno & Mayer, 2002), and, second, learners may expend unnecessary mental resources to reconcile the two sources of verbal information (Sweller, 2005). When the two sources are identical (as in the present identical-full-text and identical-abridged-text conditions), any efforts to reconcile them relies only on the surface structure; that is, learners need to make only word-level comparisons to ensure that the narration matches the on-screen text. When the two sources do not match, however (as in the present abridged-text, near-change, and far-change conditions), learners must create and compare mental representations of the verbal information to make sure the narration and the on-screen text match at a conceptual level.

The effort required to make word-level comparisons does not promote mental model construction and therefore only serves to take away from resources that could be used for learning (Sweller, 2005). Creating mental representations of the text, on the other hand, requires deeper mental processing—that is, generative, rather than extraneous, processing (Kintsch & van Dijk, 1978; Mayer, 2005). Given the striking difference in performance between our near-change and far-change conditions, we suggest that there is likely an optimal balance between redundancy and discrepancy for the promotion of learning—a slight difference between narration and on-screen text fosters generative processing within working memory limits and can thereby enhance learning, but too much difference overloads working memory and can thus prevent learners from comprehending the lesson fully. The appropriate level of correspondence may vary by the materials to be learned, the experience or background of the learner, and the pace of the lesson (Kalyuga et al., 1999). In a self-paced lesson, for example, learners may be able to take sufficient time to process the text, compare it with the narration, view the animation, and integrate all these sources of information (Betrancourt, 2005).

One reason that the additional cognitive processing induced by our abridged-text and near-change conditions remained within working memory limits could be that the on-screen motion was limited. Even though we presented an animation, only about one third of the duration of a given segment involved on-screen motion. It is possible, then, that learners had enough time to switch their attention between the images and the text without causing an overload in the visual channel (Moreno & Mayer, 1999, 2002).

The present research also offers further evidence that learners can be easily swayed or misled by feelings of fluency during learning (for a review, see Bjork, Dunlosky, & Kornell, in press). Many learners,

for example, indicated that they would have preferred a presentation in which they could see on-screen text identical to the narration at the same time as the animation. Even at the end of the experiments, when participants were allowed to type in any additional comments they had regarding the study, several in the abridged-text or near-change condition remarked that nonmatching text was “distracting” or “caused me to lose focus as I was trying to internalize two semi-conflicting messages”—even though such participants had just performed better in those conditions than they would have in a condition with identical on-screen text. These spontaneous comments, as well as participants’ responses to the metacognitive questions, clearly indicate a tendency to prefer a learning condition that appears easy at acquisition rather than one that appears more difficult but would promote better long-term retention and transfer.

Practical Implications

In light of these theoretical implications, there are several educational applications we can draw from the present research. First, people are generally not able to discern the learning situation that is likely to result in the highest recall and transfer performance based on experience alone. Consequently, both instructors and learners can be misled by feelings of fluency during learning, which can, in turn, negatively impact instructional design. It is thus important to make instructors who do use multimedia materials aware that apparent difficulties in a learning situation, such as resolving slight differences between the narration and the on-screen text, can be beneficial for learning.

Second, on-screen text should not necessarily be avoided. Previous research suggests that on-screen captions are harmful for learning when a pictorial visual aid is presented simultaneously (e.g., Kalyuga et al., 2004; Mayer et al., 2001). Our findings, however, suggest that instructors could use limited on-screen captions to support learning by highlighting key points or phrasing the point they are saying out loud in a slightly different way. The captions should be similar to the narration, but they should not be identical. It should be noted, however, that our results are based on a system-paced lesson for novices, so results may differ for self-paced or expert learners. In addition, segments of simultaneous visual and verbal information may need to be kept relatively short in order to remain within working memory capacity.

Limitations and Future Directions

The present study has some limitations. First, although the overall pattern of our results is consistent with predictions that follow from the desirable-difficulties hypothesis—namely, that the abridged-text and near-change conditions should result in the best performance, while the identical-full-text and identical-abridged conditions should result in the lowest performance—it should be noted that some of the results consistent with this conclusion were only marginally significant and, thus, need to be interpreted with caution. For example, while the advantage of the abridged-text condition over the identical-full-text condition was statistically significant on the recall test in both experiments, its advantage on the transfer tests was only marginally significant ($p = .06$, $d = 0.48$, and $p = .08$, $d = 0.32$, in Experiments 1 and 2, respectively). Nonetheless, given the moderate effect sizes and the consistent overall pattern of results between the present two experiments, we believe our interpretation to be largely supported. It

is to be hoped, however, that future research will continue to explore these same issues with the goal of furthering researchers' understanding of the processing induced by nonidentical text in multimedia presentations and the conditions under which it is beneficial versus harmful to students' retention and comprehension.

Second, we eliminated participants with too much prior knowledge of the lesson content from all of the presentation conditions, and it is possible that the effects of nonidentical text may differ for more knowledgeable learners. Indeed, greater prior expertise may have protected some learners in the identical-full-text condition in Experiment 2 from being impaired on the transfer test to the extent that learners in that condition were in Experiment 1, suggesting that how the effects of various levels of desirable difficulties in multimedia presentations might be modulated by expertise as a fruitful line for future research.

Finally, an important avenue for future research is a deeper examination of why a small discrepancy between the on-screen text and narration was beneficial while a large discrepancy was not. One possible factor to consider is that longer segments in particular might benefit from a visual presentation of the text due to the transiency of spoken information (see Leahy & Sweller, 2011). The present research suggests that short text segments that differ from a narration remain within working memory capacity, but more research is needed to clarify the exact conditions in which discrepancy benefits the integration of visual and auditory information.

Concluding Comments

For the foreseeable future, the use of multimedia materials as a primary or supplemental means of instruction can be expected to increase. It is critical, therefore, that instructors be able to design such materials in a way that blends text, narration, and on-screen animations in a maximally effective way. The present findings suggest, however, that achieving that goal rests on increased understanding of when redundancy across those components is helpful and when it is harmful; they also demonstrate that neither intuition nor prevailing practices can substitute for such an understanding.

References

- Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology, 104*, 250–263. doi:10.1037/a0026147
- Apperson, J. M., Laws, E. L., & Scepansky, J. A. (2008). An assessment of student preferences for PowerPoint presentation structure in undergraduate courses. *Computers & Education, 50*, 148–153. doi:10.1016/j.compedu.2006.04.003
- Betrancourt, M. (2005). The animation and interactivity principles in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 287–296). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819.019
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (in press). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*.
- Jamet, E., & Le Bohec, O. (2007). The effect of redundant text in multimedia instruction. *Contemporary Educational Psychology, 32*, 588–598. doi:10.1016/j.cedpsych.2006.07.001
- Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351–371. doi:10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6
- Kalyuga, S., Chandler, P., & Sweller, J. (2004). When redundant on-screen text in multimedia technical instruction can interfere with learning. *Human Factors, 46*, 567–581. doi:10.1518/hfes.46.3.567.50405
- Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469–486.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394. doi:10.1037/0033-295X.85.5.363
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science, 19*, 585–592.
- Leahy, W., Chandler, P., & Sweller, J. (2003). When auditory presentations should and should not be a component of multimedia instruction. *Applied Cognitive Psychology, 17*, 401–418. doi:10.1002/acp.877
- Leahy, W., & Sweller, J. (2011). Cognitive load theory, modality of presentation, and the transient information effect. *Applied Cognitive Psychology, 25*, 943–951. doi:10.1002/acp.1787
- Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819.004
- Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*, 187–198. doi:10.1037/0022-0663.93.1.187
- Mayer, R. E., & Johnson, C. I. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology, 100*, 380–386. doi:10.1037/0022-0663.100.2.380
- Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*, 358–368. doi:10.1037/0022-0663.91.2.358
- Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*, 156–163. doi:10.1037/0022-0663.94.1.156
- Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, England: Oxford University Press.
- Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representations. *Learning and Instruction, 13*, 141–156. doi:10.1016/S0959-4752(02)00017-8
- Sweller, J. (2005). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 159–168). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819.011
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory in perspective. In J. M. Spector & S. Lajoie, *Expectations in the learning sciences, instructional systems and performance technologies* (pp. 237–242). New York, NY: Springer. doi:10.1007/978-1-4419-8126-4_18
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. doi:10.1023/A:1022193728205

(Appendices follow)

Appendix A

Full Narration Divided Into 14 Idea Units

1. Stars are born out of nebulae, which are clouds in space made up of dust and gas. Over time, gravity causes the dust and gas to accrete and clump together to form a protostar, the very first stage in the life cycle of a star.
2. As the protostar accretes more dust and gas atoms, the density at its core increases, leading to an increase in temperature and gas pressure.
3. The star stops accreting molecules and enters the main sequence phase when it achieves equilibrium between gas pressure pushing outward and gravity pulling inward.
4. The star spends the majority of its life in this phase maintaining equilibrium.
5. Since gravity is constant and no more fuel is being pulled into the star, gas pressure must be maintained by fusing hydrogen atoms into helium atoms in the star's core. This nuclear fusion causes heat and energy to radiate into space, and the core of the star begins to heat up.
6. Once all the hydrogen in the core has been converted to helium, the star has entered old age and is called a red giant.
7. It continues to burn fuel by performing nuclear fusion, but now it must fuse helium atoms into carbon. The star is now burning fuel more rapidly and is less stable than it was in the main sequence phase.
8. As the temperature of the star increases, the outer shell of the star expands.
9. After the red giant phase, the life of a star takes a different path depending on its size. The larger a star is, the faster it must burn its fuel to maintain equilibrium and the faster it progresses through the life cycle.
10. In low-mass stars, thermonuclear explosions occur in the outer shell every few thousand years. Because of the instability these explosions produce, the outer shell of dust and gas particles expands and eventually dissipates, leaving behind only the hot core. Nuclear fusion has left the core filled with mostly carbon, which the low-mass star cannot fuse into heavier elements.
11. Without a further source of energy to create gas pressure, gravity forces the star to contract, and it remains a white dwarf for billions of years.
12. Medium-mass stars also become white dwarfs, but their life continues beyond that stage as neutron stars.
13. The greater core mass of a neutron star means that gravity is strong enough to pull in the outer layers it shed as a red giant. It continues to absorb matter until it achieves enough gas pressure to produce a powerful explosion known as a supernova.
14. The most massive stars can fuse carbon into other elements. Once they have burned through all their fuel, these stars also turn into neutron stars and produce a supernova. After the supernova, the most massive stars contain such a strong gravitational pull that they sometimes pull in even space itself. This process results in a black hole, an area in space where not even light can escape the gravitational pull.

(Appendices continue)

Appendix B

Transfer Questions and Possible Answers

1. How could a star be kept in the main sequence phase?
 - a. Replace helium atoms with two hydrogen atoms (on a larger scale)
 - b. Add hydrogen/fuel
 - c. Keep gravity at core stronger than internal pressure/keep internal pressure lower than gravity at core
2. What could prevent a medium- or high-mass star from going supernova?
 - a. An external gravitational force acts on the star (or something indicating that the student understands that the gas cloud cannot be attracted back)
 - b. Additional source of fuel
 - c. Keeping the temperature low enough
 - d. Ability to fuse heavier elements
3. What is the relationship between gas pressure and gravity in the red giant phase?
 - a. Gas pressure is higher than the force of gravity
4. What could cause two stars of the same initial mass to enter the red giant phase at different times?
 - a. Different amounts of hydrogen
 - b. External gravity affecting star
 - c. Different rate of particle accumulation or nuclear fusion
 - d. Different temperatures

Received September 7, 2012

Revision received January 14, 2013

Accepted January 17, 2013 ■

Learning With Animation and Illusions of Understanding

Eugene S. Paik and Gregory Schraw
University of Nevada, Las Vegas

The illusion of understanding hypothesis asserts that, when people are learning with multimedia presentations, the addition of animation can affect metacognitive monitoring such that they perceive the presentation to be easier to understand and develop more optimistic metacomprehension. As a result, learners invest less cognitive effort when learning with animation. This study tested the illusion of understanding hypothesis with a randomized, double-blind, 2×2 factorial design using two different types of animation—representational and directive. Representational animation had a negative effect on learning, and directive animation had a positive effect. Both representational and directive animations induced illusion of understanding. Moreover, the animations induced multiple forms of the illusion. Consistent with expertise reversal effect, the animations induced more optimistic metacomprehension in low-proficiency learners but more pessimistic metacomprehension in high-proficiency learners.

Keywords: animation, multimedia, metacognition, metacomprehension, expertise reversal effect

Supplemental materials: <http://dx.doi.org/10.1037/a0030281.supp>

Research shows that adding animation to multimedia presentations can have negative as well as positive effects on learning (Betrancourt, 2005; Betrancourt & Tversky, 2000; Höffler & Leutner, 2007; Tversky & Morrison, 2002). Multimedia theorists have postulated that animation can impede learning by perturbing metacognitive monitoring. In particular, animation can induce an *illusion of understanding* in which learners develop more optimistic metacomprehension. As a result, learners reduce their cognitive engagement when learning with animation (Betrancourt, 2005; Köhl, Scheiter, Gerjets, & Gemballa, 2011; Lewalter, 2003; Schnotz & Rasch, 2005).

Numerous studies have examined the effects of animation on learning, but few have systematically examined the effects of animation on metacognitive processes during learning. The present study fills this gap by examining claims made by the illusion of understanding hypothesis and the accuracy of performance standard hypothesis. This examination was conducted with two different types of animation—*representational* and *directive*. The results promote a clearer theoretical understanding of how animation affects learning through its influence on metacognition.

Representational and Directive Animation

Representational animation and directive animation embody two distinct cognitive strategies for helping the learner. An animation is said to be representational if it illustrates the content of

the presentation (Höffler & Leutner, 2007; Paik, 2009; Schnotz & Lowe, 2008). Typically, representational animations have been used to depict the behavior of dynamic systems as they change over time. For example, in this study representational animation was used to show the movement of the parts in a flushing toilet tank (see Figure 1). Although representational animations can explicitly portray the dynamic behavior of the system, such behaviors can only be implied with static images (Betrancourt, 2005; Schnotz & Lowe, 2003, 2008). Therefore, with static images, the learner must infer the system's behavior. In the extreme case, representational animation can enable the learner to form a mental visualization of the system's behavior that the learner is incapable of forming with only static images. Even when the learner is capable of inferring the system's behavior from static images, representational animation can facilitate the inferential process and, thereby, reduce cognitive demand (Schnotz & Lowe, 2008; Schnotz & Rasch, 2005).

An animation is said to be directive when it directs the viewer's attention to a particular component or area of an image (Schnotz & Lowe, 2008). The concept of directive animation has had numerous labels in the literature including *highlighting* (Jeung, Chandler & Sweller, 1997), *signaling* (Kriz & Hegarty, 2007), and *cuing* (de Koning, Tabbers, Rikers, & Paas, 2009). Directive animations can be implemented with a variety of techniques including flashing (i.e., quickly lightening and darkening an image area) or displaying an image area in different colors, as well as with animated pedagogical agents (e.g., cartoon characters) that point to or look in the direction of an image area (de Koning et al., 2009; Mayer, 2005; Moreno, 2005). Directive animations typically have been used to help the learner integrate aural and visual components of multimedia presentations. By synchronizing directive animation with running narration, directive animation can support the process of searching the image for the referents of the narration. In this manner, directive animation can increase the cognitive efficiency of visually identifying thematically salient components of an im-

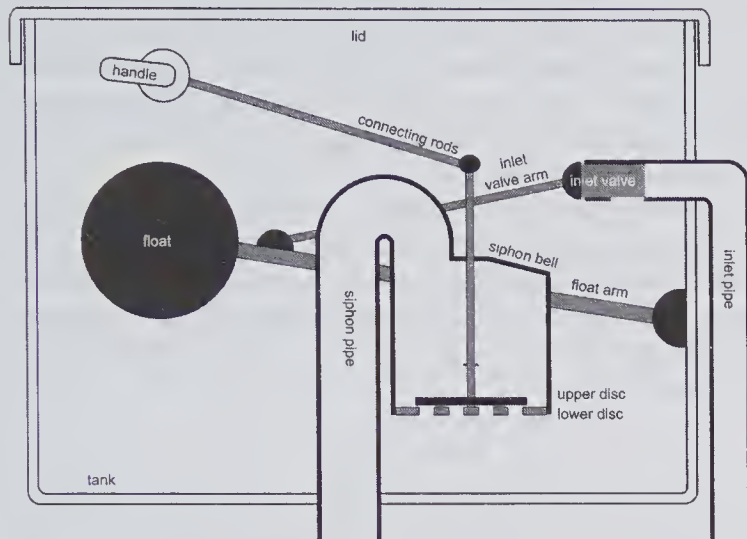
This article was published Online First October 8, 2012.

Eugene S. Paik and Gregory Schraw, Department of Educational Psychology and Higher Education, University of Nevada, Las Vegas.

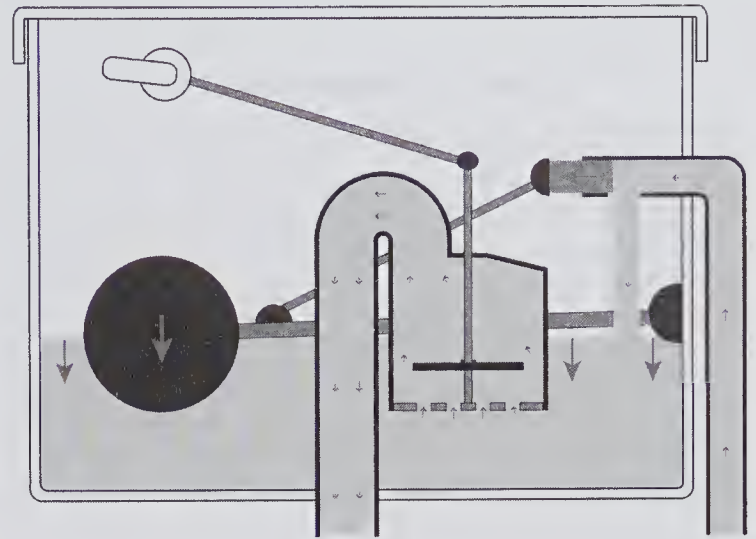
The experiment presented in this article was conducted as part of Eugene Paik's doctoral dissertation at the University of Nevada, Las Vegas.

Correspondence concerning this article should be addressed to Eugene Paik, 1943 Sunset Village Circle, Henderson, NV 89014. E-mail: paik@cox.net

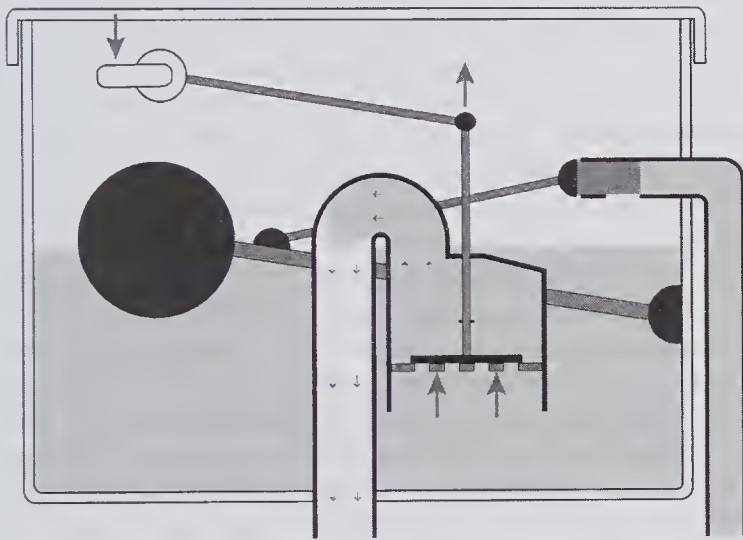
Parts of the Flushing Toilet Tank



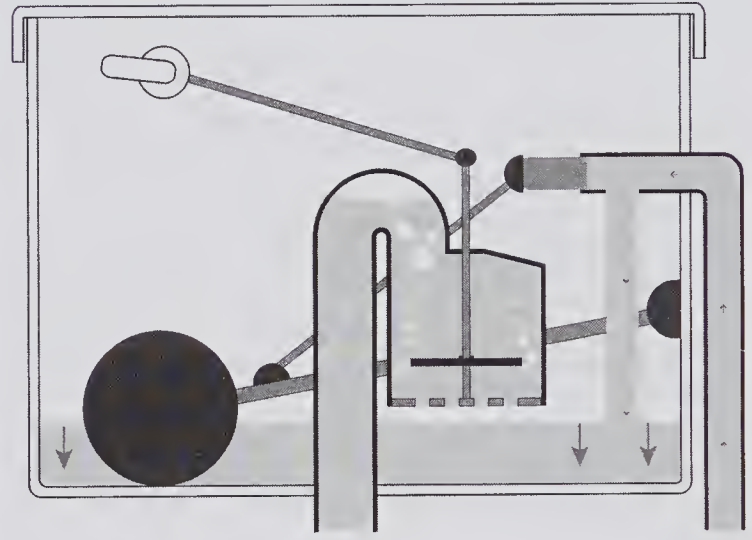
Phase 3 - Starting the Refill



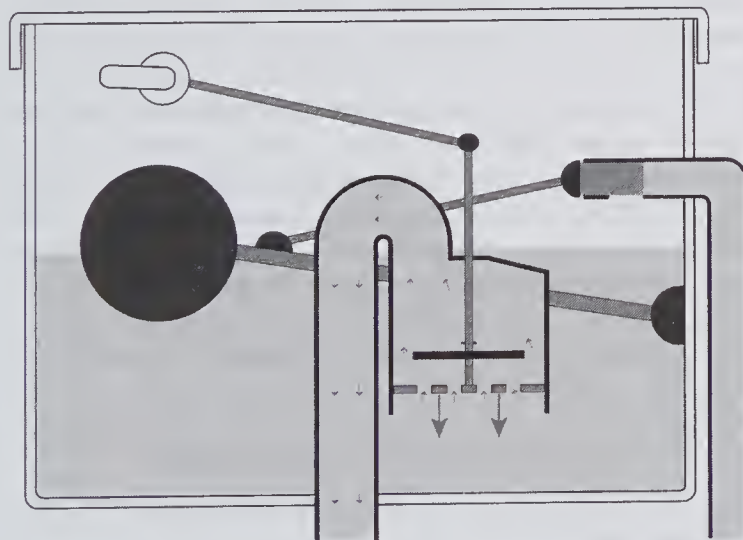
Phase 1 - Starting the Flush



Phase 4 - Ending the Flush



Phase 2 - Continuing the Flush



Phase 5 - Ending the Refill

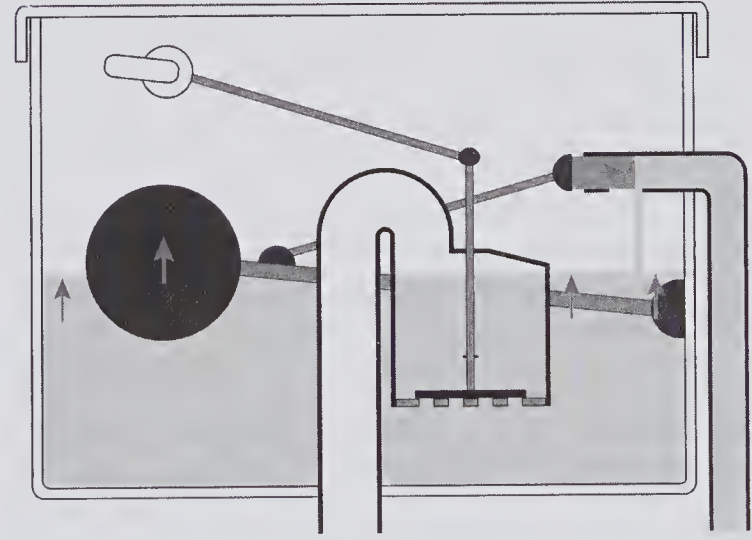


Figure 1. Six static images corresponding to the six segments of multimedia presentation on the working of a flushing toilet tank.

age, thereby enhancing learning (Jeung et al., 1997; Schnotz & Lowe, 2008).

Illusion of Understanding

The addition of directive animation in multimedia presentations generally has a positive effect on learning when the animation helps the learner integrate aural and visual components of the presentation (Höffler & Leutner, 2007). In contrast, the addition of representational animation in multimedia presentations can have negative as well as positive effects on learning (Betrancourt, 2005; Betrancourt & Tversky, 2000; Höffler & Leutner, 2007; Tversky & Morrison, 2002). Multimedia theorists have proposed a number of cognitive mechanisms by which representational animation may impede learning. For example, Hegarty, Kriz, and Cate (2003) provided some support for the idea that cognitive representation (i.e., the manner in which the behavior of dynamic systems is represented in the cognitive system) is more similar to series of static images than to animations. Mayer, Hegarty, Mayer, and Campbell (2005) provided a plausible rationale for why germane cognitive processes are engendered more by static images than by representational animations. Others argued that representational animation impedes learning because animations can negatively impact metacognition. That is, representational animations can induce an *illusion of understanding* in learners (Betrancourt, 2005; Köhl et al., 2011; Lewalter, 2003; Schnotz & Rasch, 2005).

According to the illusion of understanding (IU) model, adding animation to multimedia presentations can cause learners to overestimate how easy it is to comprehend the presented material and, thereby, develop inflated metacomprehension. As a result, learners invest less cognitive resources to the learning task. This effect is similar to the *illusion of knowing* phenomenon described in calibration research (Glenberg, Wilkinson, & Epstein, 1982; Serra & Metcalfe, 2009) in which learners overestimate the degree to which they understand information presented in text. As a result, learners allocate less attention and fail to monitor with full vigilance due to inappropriate judgments of learning.

We define the IU hypothesis as three metacognitive assertions. These assertions are based on a view of self-regulation as a dynamic cycle in which monitoring informs control processes that facilitate ongoing planning and implementation of strategies and subsequent monitoring (Azevedo & Witherspoon, 2009; Efklides, 2008; Nelson & Narens, 1990; Terricone, 2011; Winne, 2001). The *judgment of difficulty (JOD) assertion* states that, when people are learning about the behavior of dynamic systems with multimedia presentations, the addition of representational animation that explicitly illustrates the system's behavior causes them to perceive the presentation as being easier to learn (Betrancourt, 2005; Salomon, 1984; Schnotz & Rasch, 2005, 2008). With representational animation, mentally visualizing the behavior of the system is essentially a perceptual exercise. In contrast, when the behavior of the system is implied with static images, the learner must expend cognitive resources to mentally animate the system.

The *judgment of comprehension (JOC) assertion* states that the addition of representational animation causes learners to form more optimistic metacomprehension (Schnotz & Rasch, 2005). The relative ease with which learners visually experience the system's dynamic behavior with representational animation causes them to inflate their judgment of how well they comprehend that

system. Serra and Metcalfe (2009) used the term *fluency heuristics* to refer to an analogous relationship reported in text comprehension research in which metacognitive judgments of task difficulty (e.g., retrieval fluency, ease of learning) influence judgments of comprehension. Also, Serra and Dunlosky (2010) reported that beliefs about the efficacy of multimedia formats influence learners' metacomprehension judgments.

The *disengagement assertion* states that the addition of representational animation causes learners to reduce their cognitive engagement (Schnotz & Rasch, 2005). An inflated sense of comprehension leads learners to conclude that less cognitive resources and effort are needed to sufficiently comprehend the presented information.

Prior Studies

Two think-aloud studies (Köhl et al., 2011; Lewalter, 2003) provided some support for the IU hypothesis based on the frequency of comprehension and planning statements. First, in both studies animation learners uttered more positive comprehension statements than static image learners. Consistent with the JOC assertion, this difference may have resulted from animation learners having developed an inflated sense of comprehension relative to that of static image learners. Second, in the Lewalter study, animation learners uttered fewer planning statements than static image learners. Consistent with the disengagement assertion, this difference may indicate that animation learners invested less cognitive effort than static image learners in remediation.

Although the two findings described above may be consistent with animation learners having experienced an illusion of understanding, the overall pattern of data from Lewalter (2003) and Köhl et al. (2011) is also consistent with an alternate explanation, that the differences in the frequency of comprehension and planning statements were caused by differences in learning. Because the differences in learning between static image learners and animation learners were statistically not significant in the two studies, there is the possibility that animation had a positive effect on learning, but this effect was undetected due to lack of statistical power. If so, increased learning by animation learners may have been responsible for their higher frequency of comprehension statements and lower frequency of planning statements. Two additional findings in Köhl et al.'s study cast further doubt that their animation learners experienced an illusion of understanding. First, Köhl et al.'s animation learners uttered fewer erroneous statements than did static image learners, indicating that animation may have improved comprehension. Second, no significant differences in judgment of difficulty or mental effort were detected between animation and static image learners. Taken together, the results of the Lewalter study and the Köhl et al. study do not provide conclusive evidence that animation had induced illusion of understanding.

Accuracy of Performance Standard

The IU hypothesis explicates a metacognitive process by which representational animation may impede learning. We now introduce a metacognitive process by which representational animation may enhance learning. We generated the *accuracy of performance standard (APS) hypothesis* because it constitutes a plausible and theoretically important counterargument to the IU hypothesis.

According to the APS hypothesis, when the behavior of a dynamic system is explicitly illustrated with representational animation, the learner has access to a more reliable standard in memory (i.e., mental representation) by which to evaluate his or her comprehension. In contrast, when the behavior of dynamic systems is presented as a set of static images, the learner's metacomprehension must depend on the behavior of the system that the learner infers. These inferences may be inaccurate or incomplete. Therefore, the APS hypothesis asserts that, when people are learning about the behavior of dynamic systems with multimedia presentations, the addition of representational animation that explicitly illustrates the system's behavior causes them to generate more accurate JOC. This *accuracy assertion* has significant ramifications for the efficiency of the self-regulatory system because learners are better able to plan, select strategies, and allocate resources (Azevedo & Witherspoon, 2009; Nelson & Narens, 1990; Terricone, 2011; Tobias & Everson, 2009; Winne, 2001). Serra and Metcalfe (2009) referred to this sequence as the *accuracy-control link* by which accurate monitoring enhances control, which enhances self-regulation.

Present Study

This study tested the assertions of the IU and the APS hypothesis with both representational animation and directive animation. The experiment had four treatment conditions: static (i.e., no animation); representational animation only; directive animation only; and both representational and directive animations. Following the precedent of Mayer et al. (2005), two types of learning outcomes were measured. A *retention* test asked the participants to recall information explicitly provided in a multimedia presentation. A *transfer* test asked the participants to solve diagnostic and prognostic problems.

Judgment of difficulty (JOD), judgment of comprehension (JOC), and judgment of visualization (JOV) were used as dependent variables to test the assertions of the IU and APS hypotheses. JOD was estimated by asking the participants to characterize their learning experience (e.g., "How difficult was it to learn about the flushing toilet tank from the presentation?") and the multimedia presentation (e.g., "How would you rate the quality of the presentation that you just saw?"). JOC was estimated by asking the participants to predict their learning outcome. After the treatment was administered, the participants were asked to estimate how they would perform on a number of different problems. As these were the problems in the instruments used to measure their learning outcome (i.e., the retention test and the transfer test), the participants were effectively predicting their learning outcome. JOV was measured by asking the participants to mentally visualize the behavior of the flushing toilet tank that they had learned about during the treatment. They were then asked to characterize the quality of their visualization experience in terms of difficulty, accuracy, and level of detail.

The analysis utilized two covariates: spatial ability and prior knowledge. Prior studies showed that spatial ability and image type (e.g., animation vs. static images) can have interactional effects on learning (Höffler & Leutner, 2007). Prior knowledge had been shown to affect learning in general (Kalyuga, Ayres, Chandler, & Sweller, 2003).

Predictions

We did not include predictions of learning outcome because a number of potential cognitive mechanisms have been hypothesized by which animation may influence learning. As yet, however, there is an absence of a unifying framework that describes (a) the conditions under which each of these mechanisms become activated and (b) how the activated processes are integrated to produce the learning outcome. Indeed, one of our goals for this study was to contribute to the theoretical advancement with regard to (a).

The JOD assertion of the IU hypothesis was tested with the null hypothesis that representational animation learners would produce lower JOD (i.e., find the presentation easier) than no representational animation learners. The JOC assertion of the IU hypothesis was tested with the null hypothesis that representational animation learners would produce higher JOC (i.e., form more optimistic metacomprehension) than no representational animation learners. The disengagement assertion of the IU hypothesis was tested tangentially with JOV. According to the disengagement assertion, representational animation learners invest less cognitive resources and effort than no representational animation learners when learning about the behavior of a dynamic system. One cognitive activity that representational animation learners are less likely to engage in is mentally simulating the behavior of the system (Schnotz & Rasch, 2005). If so, then representational animation learners will have more difficulty mentally visualizing the system after the learning phase because they had less practice mentally visualizing the system during learning. Hence, the disengagement assertion was tested with the null hypothesis that representational animation learners would produce lower JOV (i.e., poorer quality of visualization) than no representational animation learners. The accuracy assertion of the APS hypothesis was tested with the null hypothesis that representational animation learners would generate more accurate JOC than no representational animation learners.

The predictions derived from the IU and APS hypotheses are applicable only to representational animation. They are not applicable to directive animation because the assertions of the hypotheses do not rely on the concept of directive animation. They do, however, rely on the concept of representational animation. Nevertheless, directive animation was included in the present study to test the hypothesized causal relationship between representational animation and metacognitive monitoring. Significant differences in JOD, JOC, or JOV between directive and no directive animation learners would suggest that current theories may be deficient. For example, evidence that representational and directive animations have a similar pattern of effect on metacognitive monitoring would necessitate substantive theoretical changes to account for the similarity.

Analytical Methods

Groupwise comparisons of JOD, JOC, and JOV were made not with raw scores but with *bias* scores. The bias of a variable was defined as the component of the variable that is not attributable to learning. The bias of a variable was estimated with the residual when the variable is regressed by overall learning (i.e., the average of retention and transfer test scores). This analytical approach eliminated the potential explanation that a detected difference in a variable was caused by differences in learning.

The accuracy of JOC for a group was estimated with the Pearson correlation between JOC and overall learning (Blanch-Hartigan, 2011). Groupwise comparisons of JOC accuracy were made with Fisher's Z-transformation.

Method

This study used a randomized, double-blind, 2×2 between-subjects factorial design. The two factors were the absence or presence of representational animation and the absence or presence of directive animation. This resulted in four treatment groups: static (i.e., no animation), representational animation only, directive animation only, and both animations (i.e., representational animation and directive animation).

Participants

The participants were 65 undergraduate psychology students (49 female, 16 male, $M_{\text{age}} = 24.3$ years, age range = 18 to 44 years) at a major university in the southwestern region of the United States. We do not report the participants' SAT scores because a majority of the participants did not provide them.

Material

The experiment protocol was administered by an interactive multimedia computer program that recorded all participant responses. All the participants worked on an identical model of computer hardware that included a 24-in. color monitor with 1920×1080 resolution and a headphone. A version of the computer program that was used to administer the protocol is provided as online supplemental material. With this computer program, the reader may examine the entire protocol including the instruments, the treatment (i.e., the four versions of the multimedia presentation on the workings of a flushing toilet tank), as well as the manner in which each instrument and treatment condition was administered. The following describes the components of the protocol in the order that they were administered.

Introduction. The introduction explained what was expected of the participants in the study. At the conclusion, the participants were asked to enter a *treatment code* that was provided to them by the protocol administrator.

Participant survey. The participant survey asked the participants about their demographic characteristics. The survey also asked the participants to characterize their background knowledge relevant to flushing toilet tanks.

Treatment. The treatment was a multimedia presentation on how a flushing toilet tank works that was adapted from Mayer et al. (2005) and Hegarty et al. (2003) with the addition that navigational control and the narration modality were controlled across the treatment conditions (Ginns, 2005; Low & Sweller, 2005; Mayer & Chandler, 2001; Mousavi, Low, & Sweller, 1995). Participants in this study viewed the presentation twice so as to provide a more realistic learning scenario than those in the Mayer et al. and Hegarty et al. studies.

The multimedia presentation contained six segments. Each segment contained graphic illustrations (see Figure 1) with an aural narration (see the Appendix). The first segment introduced the parts of the flushing toilet tank. The remaining segments described

the five phases of generating the flush and refilling the tank. The first segment lasted about 60 seconds, and the remaining segments lasted about 30 seconds each. The participants viewed all the segments continuously from start to finish. The participants were not provided any mechanism (e.g., pause or rewind buttons) to interrupt the presentation. At the completion of the first viewing, the presentation was paused. When the participant clicked a button on the computer screen, the presentation of the second viewing commenced.

The images in the presentation contained a limited palette of colors. The parts of the toilet tank were in shades of gray. The water and the arrows representing the flow of water were in shades of blue. The arrows that indicated how the parts of the toilet tank moved were in red.

The multimedia presentations across the four treatment conditions were identical except for their visual component. The static version presented a single still image for each segment (see Figure 1). Each static image was a key frame from the animated versions. For several static images, however, water flow arrows were added to better illustrate the directionality of the water's flow.

The directive-only version was identical to the static version except that several directive animation techniques were incorporated. In the first segment, parts appeared and disappeared from the screen in synchrony with the narration so that only those parts pertinent to the narration were displayed. For example, during the portion of the narration that stated "The flow of water into the tank is controlled by these parts . . ." only those enumerated parts were visible. Furthermore, the parts were displayed in a lighter shade until a part was explicitly referenced in the narration. At that point, the part being referenced was presented in normal shade, thereby providing the participant a clear visual indication of the part that was being referred to by the narration. In the remaining segments that described the dynamic behavior of the flushing toilet tank, blinking and tinting were used as visual cues to the narrative referents. For example, when the narration stated "When the handle is pressed down," the color of the handle was tinted red and the arrow pointing down on the handle flashed on and off.

In the representational-only version, representational animation depicted the movement of the parts and the flow of water. The representational animation and the narration were synchronized. For example, when the narration in phase 2 stated that "The two disks start to drop and separate from each other," the animation showed the two discs falling and separating from each other. The flow of water was animated by a continuous movement of arrows.

The representational-directive version incorporated both representational and directive animations described above. Logistically, we created first the representational-directive version such that directive animation and representational animation always appeared sequentially and never simultaneously. The other three versions of the presentation were then produced by replacing each unneeded animation clip with an appropriate static image.

Post-treatment survey. The post-treatment survey asked the participants to evaluate the format and the content of the multimedia presentation (e.g., "How difficult was it to learn about the flushing toilet tank from the presentation?"). The survey also described a set of problems to the participants. For each problem, the participants were asked how well they thought they would be able to solve the problem. For example, one of the questions was

as follows: "When a flushing toilet tank behaves abnormally, it is usually an indication that something in the tank is broken. How well do you think that you will be able to diagnose the cause of abnormal behaviors in flushing toilet tanks?" As the problem descriptions also described the problems used in the instruments to measure their learning outcome (i.e., the retention and transfer tests), the participants were asked, in effect, to predict their learning outcome.

Visualization exercise and survey. The visualization exercise asked the participants to mentally visualize the processes of the flushing toilet tank as described in the multimedia presentation. The participants were provided 30 seconds to perform the mental visualization. The visualization survey then asked the participants to characterize their visualization experience (e.g., "How detailed did your visualization seem to you? That is, compared to the level of detail provided in the presentation.").

Retention test. The retention test required the participants to recall information that was explicitly provided in the presentation. There were 10 *part-recall* problems and one *process-recall* problem. For each part-recall problem, an image of a part of the flushing toilet tank was displayed on the screen. The image was identical to that shown during the treatment presentation. All of the other parts were also shown, but they were visually distinguished with a lighter shade. Each part-recall problem required the participant to provide two responses: *part-name* and *part-purpose*. The participants provided their responses by selecting an entry in two drop-down list boxes. For the part-name response, the list included names that were explicitly referred to in the presentation (e.g., connecting rod, float, float arm) as well as those that were not (e.g., regulator). Similarly, the list for the part-purpose response included concepts explicitly referred to in the presentation (e.g., flow of water into the tank) as well as those that were not (e.g., flow of air into the tank). The participants were given a maximum of 20 seconds to complete each part-recall problem. If a participant did not provide both responses within 20 seconds, the participant was notified that the time limit had been exceeded and the next question was automatically displayed.

The process-recall problem asked the participants to write down all the key events of the flushing toilet tank. The participants typed their responses into a basic text editor window. To help the participants understand what was expected of them, we provided three key events as a starting point: (a) the handle is pressed; (b) float drops toward the bottom of the tank; and (c) the inlet valve is pushed in the inlet pipe. The participants were provided a maximum of 3 minutes to respond.

Transfer test. The transfer test required the participants to apply the principles that were introduced in the treatment presentation in novel situations. The transfer test consisted of two prognostic problems (e.g., "Suppose that the float were to break off from the float arm. How would the flushing toilet tank misbehave? Describe all the symptoms that you can think of.") and two diagnostic problems (e.g., "Suppose that you push down on the handle, but there is no flush. No water flows into the toilet bowl, none whatsoever. Describe all the causes that you can think of."). The participants were provided 90 seconds to complete each prognostic problem and 120 seconds to complete each diagnostic problem.

Spatial ability test. The Paper Folding Test (Ekstrom, French, Harman, & Derman, 1976) was adapted for online administration.

An earlier pilot study indicated that the Paper Folding Test posed significant cognitive demands on the participants. Therefore, only the first 10 problems of the original test were administered in the present study. In the original test, the first 10 problems were of comparable difficulty to the second 10. Also, the Paper Folding Test was administered at the end of the experiment to minimize the potential impact of cognitive fatigue induced by the test. Although this ordering left open the possibility that the treatment conditions systematically and differentially affected the participants' performance of the Paper Folding Test, we judged this more advantageous than the other order (i.e., where the differential effects of the Paper Folding Test occurs prior to the treatment) because key statistical analyses of the study did not rely on the results of the Paper Folding Test (see Results). The participants were given a maximum of 20 seconds to complete each Paper Folding Test problem. If a participant did not respond within the 20 seconds, the participant was notified that the time limit had been exceeded and the next question was automatically displayed.

Procedure

The experiment was administered in four group sessions, ranging from 10 to 22 participants each, within a span of 1 week. Each participant was seated in front of a computer. The participants were instructed to put on the headphone and were shown how to adjust the volume. They were then instructed to begin their participation by pressing a button on the screen. Upon completion of the introductory presentation, each participant had the option of either terminating participation or signing the informed consent form and continuing participation. No participant chose to terminate participation.

When the participant submitted the signed consent form, the administrator gave the participant a sheet of paper with a *treatment code*. The participant was then informed that the administrator was no longer available for assistance. Participants were instructed to proceed as best they could if any issues or questions arose during the session. Once a participant entered the treatment code, a computer program administered the experiment protocol. At the completion of the protocol, the computer program instructed the participant to notify the administrator. When the participant notified the administrator, the administrator thanked the participant and directed the participant to leave the lab.

Random assignment. Random assignment and even distribution of the participants across the treatment groups were implemented as follows. The treatment codes that were distributed to the participants were ordered such that each consecutive four treatment codes (a) contained all four treatment conditions and (b) were shuffled in random order. The treatment codes were distributed in the order that the participants submitted their signed consent forms.

Double blind control. Double blind control of the experiment was achieved as follows. First, the sheets of paper containing the treatment codes were folded and stapled so that the codes were hidden. Therefore, when the administrator handed a treatment code paper to a participant, the administrator did not know to which treatment condition the participant was being assigned. Second, the administrator did not engage in any interaction, including eye contact, with a participant once the participant had opened the treatment code paper. Third, physical partitions were erected be-

tween the participants' desks so that one participant could not see the computer screen of another participant. Finally, during the scoring process, participant responses were encoded and collated so that the scorers did not know to which individual or treatment group a response belonged.

Variables and scoring. In the descriptions that follow, "equally weighted sum" of scores indicates that the highest score in each subgroup was first scaled to 100%. The remaining scores were then scaled proportionately (i.e., without first scaling the lowest score to 0%). The following variables were employed in the present study:

- *Representational animation* indicated the absence or presence of representational animation in the treatment.
- *Directive animation* indicated the absence or presence of directive animation in the treatment.
- *Spatial ability* was the percentage of correct responses in the Paper Folding Test.
- *Prior knowledge* was the sum of the responses to questions 5 through 7 in the participant survey.
- *Judgment of difficulty (JOD)* was the sum of the responses to questions 1 through 4 in the post-treatment survey. Lower values signify greater ease and higher values signify greater difficulty. Note that although the reduction in mental effort is a key assertion of the IU hypothesis (i.e., the disengagement assertion), the survey item on mental effort (i.e., "How much mental effort was required to learn about the flushing toilet tank from the presentation?") was aggregated into JOD due to challenges of developing a reliable instrument for mental effort.
- *Judgment of comprehension (JOC)* was the sum of the responses to questions 5 through 12 in the post-treatment survey.
- *Judgment of visualization (JOV)* was the sum of the responses to the three questions in the visualization survey.
- *Retention* was the equally weighted sum of the part-recall score and the process-recall score from the retention test. For the part-recall score, one point was awarded for each correct part-name response and one point was awarded for each correct part-purpose response. The procedure for scoring the process-recall problem was adapted from Hegarty et al. (2003). A key was developed that consisted of 24 distinct processes of the flushing toilet tank that were explicitly described in the presentation (e.g., "the connecting rod rises," "the connecting rod pulls up the lower disk"). One point was awarded for each key process that the

participant mentioned in his or her response. Fictitious or incorrect responses were ignored. The order of the processes was also ignored.

- *Transfer* was the equally weighted sum of the scores of the four problems in the transfer test. The transfer problems were scored based on a predefined key in a manner similar to the process-recall problem.

- *Overall learning* was the standardized equally weighted sum of retention and transfer.

Results

Scores

The five free-response problems (i.e., the process-recall problem in the retention test and the four problems in the transfer test) were scored by two individuals. The Pearson correlation between the two scorers was .87. All the discrepancies between the two scorers were resolved through discussion and consensus. JOC and JOV were transformed by $x^{3/2}$ to reduce their skewness. After the transformations, the Cronbach's alpha for JOD, JOC, and JOV were .77, .88, and .87, respectively. Table 1 provides the descriptive statistics of the experimental variables. Table 2 provides the correlations between the variables.

Learning Outcomes

A 2 × 2 (representational animation × directive animation) factorial analysis of covariance (ANCOVA) was conducted using retention and transfer as dependent variables and spatial ability as a covariate. Prior knowledge was not used as a covariate because it was not significantly correlated with either dependent variable. Levene's homogeneity of variance test was not significant in the analyses.

For retention, there were no significant main effects. Spatial ability had a positive effect on retention, $F(1, 64) = 18.4, p < .001, \eta^2 = .23$.

For transfer, main effects occurred with representational animation and with directive animation. Representational animation learners performed moderately worse on the transfer test than no-representational animation learners, $F(1, 64) = 5.38, p = .024, \eta^2 = .082$. Also, directive animation learners performed moderately better on the transfer test than no-directive animation learn-

Table 1
Descriptive Statistics of Experimental Variables by Treatment Condition

Variable	Treatment condition				Total M (SD)
	Static M (SD)	Representative only M (SD)	Directive only M (SD)	Representative and directive M (SD)	
N	17	16	16	16	65
Prior knowledge	10.4 (4.45)	12.1 (3.30)	9.94 (3.07)	9.19 (3.87)	10.4 (3.78)
Spatial ability	.571 (.193)	.556 (.163)	.544 (.163)	.506 (.224)	.545 (.185)
Retention	.537 (.239)	.529 (.189)	.549 (.153)	.471 (.181)	.522 (.192)
Transfer	.421 (.166)	.359 (.184)	.531 (.181)	.394 (.131)	.426 (.175)
Overall learning	.48 (.188)	.44 (.16)	.54 (.14)	.43 (.15)	.47 (.16)
JOD	11.24 (7.16)	9.00 (3.86)	9.69 (3.72)	9.69 (4.14)	9.92 (4.93)
JOC	341.2 (146.3)	369.3 (99.1)	347.7 (73.9)	350.1 (78.0)	351.9 (102.3)
JOV	103.3 (29.5)	95.7 (30.2)	103.8 (20.1)	88.1 (30.5)	97.8 (28.0)

Note. SD = standard deviation; JOD = judgment of difficulty; JOC = judgment of comprehension; JOV = judgment of visualization.

Table 2
Pearson Correlation Between Experimental Variables

Variable	Directive	Prior knowledge	Spatial ability	JOD	JOC	JOV	Recall	Transfer	Overall learning
Representational	.015	.059	-.072	-.117	.075	-.209	-.111	-.280*	-.215
Directive		-.220	-.105	-.047	-.029	-.066	-.061	.207	.074
Prior knowledge			.158	-.335**	.474**	.430**	.232	.001	.140
Spatial ability				-.238	.258*	.300*	.491**	.323**	.457**
JOD					-.768**	-.609**	-.516**	-.372**	-.500**
JOC						.735**	.625**	.452**	.604**
JOV							.624**	.436**	.598**
Recall								.610**	.908**
Transfer									.885**

Note. JOD = judgment of difficulty; JOC = judgment of comprehension; JOV = judgment of visualization.

* $p < .05$. ** $p < .01$.

ers, $F(1, 64) = 4.64$, $p = .035$, $\eta^2 = .072$. Representational animation \times directive animation interaction was not a significant predictor of retention or of transfer.

Illusion of Understanding

We regressed representational animation, overall learning, and representational animation \times overall learning (henceforth abbreviated as $R \times L$) on dependent variables JOD, JOC, and JOV in three independent analyses. Also, the three regression analyses were replicated with directive animation in place of representational animation. These regression analyses provided a comparison of the dependent variable among participants with *comparable learning*, so that any detected difference in the dependent variable may not be attributed to variance in learning. When a significant interaction was found, we calculated the intersection point between the two regression lines (one for the absence and the other for the presence of respective animation). We then calculated the region of significance using the Johnson–Neyman method (Preacher, Curran, & Bauer, 2006). The results of these regression analyses are provided in Table 3. How these analyses support and refute of the assertions of the IU hypothesis is summarized in Table 4.

Representational animation. The effect of representational animation on JOD was consistent with the JOD assertion, and the effect of representational animation on JOC was consistent with the JOC assertion. However, the effects were consistent only for a subgroup of participants, namely, those with low overall learning scores. First, $R \times L$ was a significant predictor of JOD and of JOC, indicating that the effects of representational animation on JOD and on JOC were moderated by overall learning. Second, for both JOD and JOC, the $R \times L$ interaction was disordinal. That is, the regression lines intersected within the region of interest (i.e., overall learning within $M \pm 2 SD$). Third, the region of significance analysis revealed that, for overall learning less than $-.070 SD$, representational animation learners developed lower JOD (i.e., found the presentation easier to understand) than no-representational animation learners. For overall learning less than $-.058 SD$, representational animation learners produced higher JOC (i.e., more optimistic metacomprehension) than no-representational animation learners.

JOV was not significantly affected by representational animation or by $R \times L$. Therefore, the IU disengagement assertion was neither supported nor refuted by the analysis.

Directive animation. Contrary to the predictions of the IU hypothesis (i.e., directive animation would have no effect on JOD, JOC, and JOV), directive animation had a number of significant effects on JOD and on JOC. Further surprisingly, directive animation and representational animation had a strikingly similar pattern of effect on JOD, JOC, and JOV. First, the $D \times L$ interaction was a significant predictor of JOD and of JOC. Second, for both JOD and JOC, the $D \times L$ interaction was disordinal. Third, the lower bounds of the regions of significance were within the regions of interest (i.e., $-.519 SD$ for JOD; $-1.84 SD$ for JOC). Finally, JOV was not significantly affected by directive animation or by $D \times L$.

Accuracy of Performance Standard

The accuracy of JOC for a group was estimated with the correlation of JOC and overall learning (referred to below as the JOC/L correlation) within the group. The JOC accuracy of two groups was compared by comparing the JOC/L correlations of the two groups with Fisher's Z-transformation.

The APS assertion (i.e., representational animation would produce more accurate JOC) was neither supported nor refuted by the analysis. For representational animation learners, the JOC/L correlation was $r = .500$, $N = .32$, $p = .004$. For no-representational animation learners, the correlation was $r = .734$, $N = .33$, $p < .001$. However, the difference between the two correlations was not statistically significant.

With respect to directive animation, the results contradicted the predictions of the APS hypothesis (i.e., directive animation would have no effect on JOC accuracy). Directive animation learners produced significantly less accurate JOC than no-directive animation learners. For directive animation learners, the JOC/L correlation was $r = .395$, $N = .32$, $p = .025$. For no-directive animation learners, the correlation was $r = .730$, $N = .33$, $p < .001$. The difference between the two correlations was significant, $Z = 1.857$, $p = .05$.

Static Versus Animation Groups

We also compared JOC accuracy between participants whose multimedia presentation did not contain any form of animation (i.e., the static image group) and those whose multimedia presentation contained some form of animation (i.e., representational-

Table 3
Regression Statistics for JOD, JOC, and JOV

Dependent variable factor	β	B	95% CI of B		t	p^a	Intersection point	Significance region ^b	
			LL	UL				LL	UL
Representational animation									
JOD							0.799	0.070	broi
R	−.228	−2.226	−4.250	−0.202	−2.200	.032			
L	−.803	−3.956	−5.326	−2.586	−5.774	.000			
R × L	.374	2.786	0.736	4.837	2.717	.009			
JOC							1.214	0.058	broi
R	.211	42.827	2.867	82.787	2.143	.036			
L	.804	82.322	55.269	109.375	6.085	.000			
R × L	−.228	−35.282	−75.766	5.202	−1.743	.086			
JOV							n/a	n/a	n/a
R	−.084	−4.677	−16.305	6.950	−0.804	.424			
L	.563	15.785	7.913	23.657	4.010	.000			
R × L	.024	1.036	−10.744	12.816	0.176	.861			
Directive animation									
JOD							0.046	−0.519	.668
D	−.015	−0.143	−2.176	1.890	−0.141	.889			
L	−.764	−3.764	−5.113	−2.414	−5.578	.000			
D × L	.406	3.081	1.008	5.154	2.972	.004			
JOC							−0.270	−1.842	.580
D	−.071	−14.353	−53.510	24.803	−0.733	.466			
L	.830	84.941	58.958	110.924	6.537	.000			
D × L	−.338	−53.258	−93.181	−13.336	−2.668	.010			
JOV							n/a	n/a	n/a
D	−.110	−6.111	−17.418	5.196	−1.081	.284			
L	.657	18.402	10.900	25.905	4.905	.000			
D × L	−.078	−3.345	−14.873	8.183	−0.580	.564			

Note. CI = confidence interval; LL = lower limit; UL = upper limit; JOD = judgment of difficulty; JOC = judgment of comprehension; JOV = judgment of visualization; R = representative; D = directive; L = overall learning; broi = beyond region of interest.
^a.000 means $p < .0005$. ^bRegions below lower bound and above upper bound are significant at $\alpha = .05$.

only, directive-only, or representational-directive). Animation learners produced significantly less accurate JOC than static image learners. For animation learners, the JOC/L correlation was $r = .445, N = .48, p < .002$. For static image learners, the correlation was $r = .856, N = .17, p < .001$. The difference between the two correlations was significant, $Z = 2.613, p = .009$.

Expertise Reversal Effect

The disordinal interactions of R \times L and of D \times L with respect to JOD and JOC indicated that the animations may exhibit expertise reversal effect. That is, the animations may have opposite effects on low- and high-proficiency learners. With representa-

Table 4
Confirmations and Refutations of IU and APS Assertions

Hypothesis assertion	Measure	Representational animation		Directive animation	
		Prediction	Result	Prediction	Result
IU hypothesis					
JOD assertion	JOD bias	R < noR	Confirmed: L < .070	D = noD	Refuted: D < noD, where L < -.519 D > noD, where L > .668
JOC assertion	JOC bias	R > noR	Confirmed: L < .058	D = noD	Refuted: D > noD, where L < -1.842 D < noD, where L > .580
Disengagement assertion	JOV bias	R < noR	n.s.	D = noD	n.s.
APS hypothesis					
Accuracy assertion	JOC accuracy	R > noR	n.s.	D = noD	Refuted: D < noD

Note. IU = illusion of understanding; APS = accuracy of performance standard; JOD = judgment of difficulty; JOC = judgment of comprehension; JOV = judgment of visualization; R = Presence of representational animation; noR = Absence of representational animation; D = Presence of directive animation; noD = Absence of directive animation; L = overall learning (the units of overall learning are standard deviations); n.s. = not significant.

tional animation, only the lower bounds of the regions of significance were within the regions of interest. Therefore, only one group of participants (i.e., the low-proficiency learners) produced statistically significant differences in JOD and JOC between representational animation and no-representational animation learners.

With directive animation, however, both the lower bounds and the upper bounds of the regions of significance were within the regions of interest. Therefore, two groups of participants (i.e., the low-proficiency learners and the high-proficiency learners) produced statistically significant differences in JOD and JOC between directive animation and no-directive animation learners. Consistent with the expertise reversal effect, the two groups experienced opposite forms of illusion of understanding. Low-proficiency learners experienced *optimistic* illusion of understanding, where animation learners find the presentation easier to understand and inflate their metacomprehension. High-proficiency learners experienced *pessimistic* illusion of understanding, where animation learners find the presentation more difficult to understand and deflate their metacomprehension.

Discussion

This study examined the effects of representational and directive animations on learning and metacognition. We tested the illusion of understanding (IU) hypothesis, which predicted that, when people are learning about the behavior of a dynamic system with multimedia presentations, the addition of representational animation that explicitly illustrates the system's behavior would cause them to (a) find the presentations easier to understand, (b) inflate their metacomprehension, and (c) produce poorer mental visualizations of the system. We also tested the accuracy of performance standard (APS) hypothesis, which predicted that adding representational animation would enhance the accuracy of learners' metacomprehension. Both the IU and the APS hypothesis predicted that directive animation would have no effect on metacognitive monitoring.

The results indicated that representational animation induced an illusion of understanding (i.e., decreased JOD and increased JOC) as predicted by the IU hypothesis; however, this illusion was induced only for low-proficiency learners (i.e., learners who had low overall learning scores). Representational animation did not have a statistically significant effect on JOC accuracy. Overall, the pattern of effect of representational animation and of directive animation on JOD and JOC indicated that adding animation to multimedia presentations affects metacognitive monitoring in complex ways that were not predicted by the IU or the APS hypothesis.

Learning Outcome

The negative effects of representational animation and the positive effects of directive animation on transfer learning found in the present study are consistent with prior studies (de Koning et al., 2009; Hegarty et al., 2003; Höffler & Leutner, 2007; Mayer et al., 2005; Schnotz & Lowe, 2005). The treatment conditions and the outcome measures used in this study were adapted from Experiment 2 of Mayer et al. In that study, the effect size of representational animation on transfer learning

was Cohen's $d = 1.06$. In the present study, the effect size was $d = 0.56$.

Illusion of Understanding

With respect to representational animation, the assertions of the IU hypothesis were partially supported. Representational animation induced an illusion of understanding (Betrandcourt, 2005; Kühl et al., 2011; Lewalter, 2003; Schnotz & Rasch, 2005) but only among a subpopulation of participants with low overall learning scores. We refer to these learners as low-proficiency learners. Among low-proficiency learners, representational animation learners found the multimedia presentation easier and produced more optimistic metacomprehension. That is to say, low-proficiency representational animation learners experienced an illusion of understanding.

On the other hand, the overall pattern of effect of representational animation and of directive animation on JOD and JOC was more complex than predicted by the IU hypothesis. First, the effects of representational animation and of directive animation were moderated by learning proficiency. That is, there were attribute-treatment interactions (Cronbach & Snow, 1977) of $R \times L$ and of $D \times L$ on JOD and JOC. Second, the effects of directive animation were strikingly similar to the effects of representational animation. Indeed, the effects of directive animation were even more pronounced than those of representational animation. Directive animation induced two different forms of IU. *Optimistic* IU refers to lower JOD and higher JOC. *Pessimistic* IU refers to higher JOD and lower JOC. Directive animation induced optimistic IU in low-proficiency learners but pessimistic IU in high-proficiency learners. This bipolar effect of directive animation was consistent with expertise reversal effect (Plass, Kalyuga, & Leutner, 2010).

Accuracy of Performance Standard

With respect to representational animation, the results neither refuted nor supported the JOC accuracy assertion that representational animation enhances JOC accuracy. The difference in JOC accuracy between representational animation learners and no representational animation learners was not statistically significant. With respect to directive animation, the results contradicted the prediction of the APS hypothesis that directive animation would not affect JOC accuracy. Directive animation significantly degraded JOC accuracy.

Conclusion

We began this paper with the observation that representational animation has bipolar effects on learning. That is, representational animation can have a positive effect on learning under certain conditions but a negative effect under different conditions (Betrandcourt, 2005; Höffler & Leutner, 2007; Tversky, & Morrison, 2002). Multimedia theorists offered the IU model as a metacognitive explanation for why representational animation has sometimes negatively affected learning (Betrandcourt, 2005; Schnotz & Rasch, 2005). The results of this study showed that the IU model can explain not only the negative effects of representational animation but also the bipolar effects

of animation in general. Learning is impeded when animation induces optimistic IU, and learning is enhanced when animation induces pessimistic IU.

Although the results showed that animation perturbs meta-cognitive monitoring (e.g., raises and lowers JOD and JOC), a causal explanation for this perturbation remains unresolved. Explanations that rely on the representational characteristics of animation (e.g., the IU and the APS hypotheses as defined in the introduction) do not explain the following: (a) Why were the effects of animation moderated by learning? (b) Why did directive animation affect JOD and JOC? (c) Why were the effects of representational and directive animations so similar to each other? However, given that the present study included one experiment using one set of content material (i.e., the workings of a flushing toilet tank), our findings should be replicated and extended to new materials.

References

- Azevedo, R., & Witherspoon, A. M. (2009). Self-regulated use of hypermedia. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 319–339). Mahwah, NJ: Erlbaum.
- Betrancourt, M. (2005). The animation and interactivity principle of multimedia learning. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 287–296). New York, NY: Cambridge University Press.
- Betrancourt, M., & Tversky, B. (2000). Effects of computer animation on users' performance: A review. *Le Travail Humain*, 63, 311–329.
- Blanch-Hartigan, D. (2011). Medical students' self-assessment of performance: Results from three meta-analyses. *Patient Education and Counseling*, 84, 3–9. doi:10.1016/j.pec.2010.06.037
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interaction*. New York, NY: Irvington.
- de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2009). Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review*, 21, 113–140. doi:10.1007/s10648-009-9098-7
- Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, 13, 277–287. doi:10.1027/1016-9040.13.4.277
- Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.
- Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, 15, 313–331. doi:10.1016/j.learninstruc.2005.07.001
- Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, 10, 597–602. doi:10.3758/BF03202442
- Hegarty, M., Kriz, S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition and Instruction*, 21, 209–249. doi:10.1207/s1532690xci2104_1
- Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, 17, 722–738. doi:10.1016/j.learninstruc.2007.09.013
- Höffler, T. N., & Leutner, D. (2011). The role of spatial ability in learning from instructional animations—Evidence for an ability-as-compensator hypothesis. *Computers in Human Behavior*, 27, 209–216. doi:10.1016/j.chb.2010.07.042
- Jeung, H., Chandler, P., & Sweller, J. (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology*, 17, 329–345. doi:10.1080/0144341970170307
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38, 23–31. doi:10.1207/S15326985EP3801_4
- Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human-Computer Studies*, 65, 911–930. doi:10.1016/j.ijhcs.2007.06.005
- Kühl, T., Scheiter, K., Gerjets, P., & Gemballa, S. (2011). Can differences in learning strategies explain the benefits of learning from static and dynamic visualizations? *Computers & Education*, 56, 176–187. doi:10.1016/j.compedu.2010.08.008
- Lewalter, D. (2003). Cognitive strategies for learning from static and dynamic visuals. *Learning and Instruction*, 13, 177–189. doi:10.1016/S0959-4752(02)00019-1
- Low, R., & Sweller, J. (2005). The modality principle in multimedia learning. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 147–158). New York, NY: Cambridge University Press.
- Mayer, R. E. (2005). Principles of multimedia learning based on social cues: Personalization, voice, and image principles. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 201–212). New York, NY: Cambridge University Press.
- Mayer, R. E., & Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of Educational Psychology*, 93, 390–397. doi:10.1037/0022-0663.93.2.390
- Mayer, R. E., Hegarty, M., Mayer, S., & Campbell, J. (2005). When static media promote active learning: Annotated illustrations versus narrated animation in multimedia instruction. *Journal of Experimental Psychology: Applied*, 11, 256–265. doi:10.1037/1076-898X.11.4.256
- Moreno, R. (2005). Multimedia learning with animated pedagogical agents. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 507–524). New York, NY: Cambridge University Press.
- Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, 87, 319–334. doi:10.1037/0022-0663.87.2.319
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173.
- Paik, E. (2009). Learning about dynamic systems with multimedia presentations containing motion animation and highlighting animation. In G. Siemens & C. Fulford (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2009* (pp. 684–693). Chesapeake, VA: AACE.
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31, 437–448. doi:10.3102/10769986031004437
- Salomon, G. (1984). Television is “easy” and print is “tough”: The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology*, 76, 647–658. doi:10.1037/0022-0663.76.4.647
- Schnotz, W., & Lowe, R. (2003). External and internal representations in multimedia learning. *Learning and Instruction*, 13, 117–123. doi:10.1016/S0959-4752(02)00015-4
- Schnotz, W., & Lowe, R. K. (2008). A unified view of learning from animated and static graphics. In R. K. Lowe & W. Schnotz (Eds.), *Learning with animation: Research implications for design* (pp. 304–356). New York, NY: Cambridge University Press.
- Schnotz, W., & Rasch, T. (2005). Enabling, facilitating, and inhibiting effects of animation in multimedia learning: Why reduction of cognitive load can have negative results on learning. *Educational Technology Research and Development*, 53, 47–58. doi:10.1007/BF02504797
- Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory*, 18, 698–711. doi:10.1080/09658211.2010.506441
- Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 278–298). Mahwah, NJ: Erlbaum.
- Sweller, J. (2005). Implications of cognitive load theory for multimedia

- learning. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 19–30). New York, NY: Cambridge University Press.
- Terricone, P. (2011). *The taxonomy of metacognition*. Hove, England: Psychology Press.
- Tobias, S., & Everson, H. T. (2009). The importance of knowing what you know. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 107–127). Mahwah, NJ: Erlbaum.
- Tversky, B., & Morrison, J. B. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, 57, 247–262. doi: 10.1006/ijhc.2002.1017
- Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153–189). Mahwah, NJ: Erlbaum.

Appendix

Treatment Narration

The narration of the six segments of the multimedia presentation in the treatment was as follows:

1. A flushing toilet tank is made up of a number of parts. The tank and the lid store the water used to flush the toilet and house the other parts. The flow of water into the tank is controlled by these parts: the inlet pipe, the inlet valve, the inlet valve arm, the float arm, and the float. The rise and fall of the float pushes the inlet valve in and out of the inlet pipe. The flow of water out of the tank and into the toilet bowl is controlled by these parts: the handle, the connecting rods, the upper disc, the lower disc, the siphon bell, and the siphon pipe. The upper disc is free to move up and down on its own. The lower disc can be moved up and down with the handle.

2. Phase 1 - Starting the flush. When the handle is pressed down, the connecting rods are pulled up, causing the lower disk to rise and to push up the upper disk. As a result, the water in the siphon bell is forced over the siphon pipe into the toilet bowl.

3. Phase 2 - Continuing the flush. Once the handle is released, the two disks start to drop and separate from each other. As a result, the water flows through the holes in the lower disk, around the edges of the upper disk, over the siphon pipe, and into the toilet

bowl. Note: the two disks separate, because the water that flows through the holes in the lower disk pushes up the upper disk.

4. Phase 3 - Starting the refill. As the water flows out of the tank, the water level drops. As the water level falls, the float drops toward the bottom, pulling out the inlet valve, and uncovering the hole in the inlet pipe. This allows the water to flow into the tank.

5. Phase 4 - Ending the flush. When the water flows out of the tank as well as into the tank, the water level continues to drop because the flow of water out of the tank is faster than into the tank. When the water level falls below the bottom of the siphon bell, air enters and breaks the siphon. This stops the flow of water into the toilet bowl.

6. Phase 5 - Ending the refill. When water flows into the tank but not out of the tank, the water level rises. As the water level rises, the float rises, pushing in the inlet valve, and closing the inlet hole. When the water level rises high enough, the flow of water into the tank is stopped. Now, the tank is ready for the next flush.

Received October 16, 2011

Revision received August 22, 2012

Accepted September 4, 2012 ■

Explanation Feedback Is Better Than Correct Answer Feedback for Promoting Transfer of Learning

Andrew C. Butler
Duke University

Namrata Godbole
University of North Carolina–Greensboro

Elizabeth J. Marsh
Duke University

Among the many factors that influence the efficacy of feedback on learning, the information contained in the feedback message is arguably the most important. One common assumption is that there is a benefit to increasing the complexity of the feedback message beyond providing the correct answer. Surprisingly, studies that have manipulated the content of the feedback message in order to isolate the unique effect of greater complexity have failed to support this assumption. However, the final test in most of these studies consisted of a repetition of the same questions from the initial test. The present research investigated whether feedback that provides an explanation of the correct answer promotes superior transfer of learning to new questions. In 2 experiments, subjects studied prose passages and then took an initial short-answer test on concepts from the text. After each question, they received correct answer feedback, explanation feedback, or no feedback (Experiment 1 only). Two days later, subjects returned for a final test that consisted of both repeated questions and new inference questions. The results showed that correct answer feedback and explanation feedback led to equivalent performance on the repeated questions, but explanation feedback produced superior performance on the new inference questions.

Keywords: feedback, learning, retention, transfer

Feedback is a critical component of any learning process because it allows learners to reduce the discrepancy between actual and desired knowledge (Black & Wiliam, 1998). Although prior research has identified many factors that influence the efficacy of feedback (for reviews, see D. L. Butler & Winne, 1995; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008), the content of the feedback message is arguably the most important aspect of any feedback procedure. The information supplied in the feedback message is critical because it enables learners to correct errors (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005) and maintain correct responses (e.g., A. C. Butler, Karpicke, & Roediger, 2008). Thus, a primary objective of feedback research is to determine what information the feedback message should contain in order to be maximally effective.

On the basis of a wealth of studies in the literature, the current answer is that the feedback message should contain the correct answer. At the most basic level, feedback must convey information about the veracity of the learner's response (i.e., correct vs. incorrect). However, many studies have shown little or no benefit of providing verification feedback relative to no feedback (e.g., Pashler et al., 2005; Plowman & Stroud, 1942; Roper, 1977; but see Fazio, Huelser, Johnson, & Marsh, 2010). Including the correct answer in the feedback message substantially increases the efficacy of feedback because it provides the information that learners need to correct their errors. Indeed, the vast majority of studies that have compared correct answer feedback with verification feedback have shown a superiority of correct answer feedback (e.g., Pashler et al., 2005; Phye & Sanders, 1994; Roper, 1977; Travers, Van Wagenen, Haygood, & McCormick, 1964; Whyte, Karolick, Neilsen, Elder, & Hawley, 1995).

Is it beneficial for the feedback message to include other information in addition to the correct answer? A common assumption among educators and researchers is that providing students with additional information in the feedback message will improve learning. The umbrella term *elaborative feedback* is often used to describe any type of feedback that is more complex than correct answer feedback, and there are many ways of elaborating the feedback message (for a taxonomy of feedback messages, see Kulhavy & Stock, 1989). Examples of elaborative feedback include providing an explanation of why a particular response is correct or incorrect (*explanation feedback*) and re-presenting the original learning materials (*restudy feedback*). Due to the assump-

This article was published Online First December 17, 2012.

Andrew C. Butler, Psychology and Neuroscience, Duke University; Namrata Godbole, Department of Psychology, University of North Carolina–Greensboro; Elizabeth J. Marsh, Psychology and Neuroscience, Duke University.

This research was supported by a Collaborative Activity Award from the James S. McDonnell Foundation's 21st Century Science Initiative in Bridging Brain, Mind and Behavior awarded to Elizabeth J. Marsh. We thank Katherine Rawson and the members of the Marsh Lab for their helpful comments and suggestions on drafts of the article.

Correspondence concerning this article should be addressed to Andrew C. Butler, Psychology and Neuroscience, Duke University, Box 90086, Durham, NC 27708-0086. E-mail: andrew.butler@duke.edu

tion that elaborative feedback is helpful to students, it is often included as a component in methods of instruction, such as intelligent tutoring systems (Corbett, Koedinger, & Anderson, 1997) and computer-assisted instruction programs (Gibbons & Fairweather, 1998). For instance, the AutoTutor is an intelligent tutoring system that helps learners solve complex physics problems by providing many different types of feedback, including hints, corrections, and explanations (Graesser, Chipman, Haynes, & Olney, 2005). However, elaborative feedback is just one of many components combined to enhance student learning in such systems, and its independent contribution to learning is not assessed.

Surprisingly, studies that have directly compared elaborative feedback with correct answer feedback have found little or no benefit to increasing the complexity of the feedback message (for a review, see Kulhavy & Stock, 1989; for a meta-analysis, see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). For example, many studies have found that there is no benefit of providing explanation feedback relative to correct answer feedback (e.g., Gilman, 1969; Kulhavy, White, Topp, Chan, & Adams, 1985; Mandernach, 2005; Pridemore & Klein, 1995; Sassenrath & Gaverick, 1965; Smits, Boon, Sluijsmans, & van Gog, 2008; Whyte et al., 1995). Similarly, other studies have shown that providing restudy feedback yields equivalent performance to correct answer feedback (e.g., Andre & Thieman, 1988; Kulhavy et al., 1985; Peeck, 1979). Critically, the content of the feedback message was manipulated as an independent variable in these studies, which allowed the unique effect of greater complexity (or lack thereof) to be isolated.

The lack of empirical support for the efficacy of elaborating the feedback message is surprising, but these null effects may be due to how learning was assessed on the final test. Almost all of the studies on elaborative feedback have used a final test that assessed retention of the correct answer by repeating the same questions from the initial test. If the learner only needs to remember the correct answer to perform well on the final test, then the additional information contained in elaborative feedback is superfluous. However, this additional information may be important for fostering better comprehension of the material. For example, providing an explanation of why a response is correct (i.e., explanation feedback) might help the learner to move from superficial factual knowledge to a more complex understanding of the concept. Thus, elaborative feedback might be expected to facilitate performance on a final test that assesses understanding rather than retention of the correct response. One hallmark of superior understanding is the ability to transfer knowledge to new contexts. *Transfer* can be broadly defined as "the influence of prior learning (retained until the present) upon the learning of, or response to, new material . . ." (McGeoch, 1942, p. 394). In the present study, we assess understanding by investigating learners' ability to transfer their knowledge on a final test that involves making inferences using previously learned concepts.

Experiment 1

The goal of the first experiment was to investigate the hypothesis that the efficacy of elaborative feedback depends on how learning is assessed. Subjects studied a set of passages and then took an initial test on critical concepts from the passages. After each question, they received explanation feedback, cor-

rect answer feedback, or no feedback. Two days later, they took a final test that assessed both retention (via repeated questions from the initial test) and transfer (via new inference questions). We predicted that explanation feedback would lead to better transfer relative to correct answer feedback, but the two types of feedback would produce equivalent retention of the correct answers.

Method

Participants. Sixty Duke University students participated for either course credit or payment. Four additional subjects were excluded because they failed to follow experimental instructions.

Design. The experiment had a 3 (type of feedback: no feedback, correct answer, explanation) \times 2 (type of final test question: repeated, new) mixed factorial design. Type of feedback was manipulated between subjects, and type of final test question was manipulated within subjects, between materials.

Materials and counterbalancing. Materials consisted of 10 passages about a variety of topics (e.g., the respiratory system, tropical cyclones, etc.) and associated questions. Six of the passages and the associated questions were adapted from A. C. Butler (2010), and the rest were created to match. Each passage consisted of 500 words of text and contained two critical concepts (see Appendix A for sample passages). Thus, there was a total of 20 critical concepts. A *concept* was operationally defined as a piece of information that must be abstracted from multiple sentences. Two questions were associated with each concept: a definition question and an inference question. All of the definition questions were used on the initial test to assess memory for the 20 concepts. The definition question was repeated on the final test for 10 of the concepts in order to assess retention of the correct answer. The inference question was given on the final test for the other 10 concepts in order to assess transfer of knowledge. The materials were counterbalanced by creating two versions of the final test. In each version, one of the two concepts for each passage was tested by repeating the definition question, whereas the other concept was tested with a new inference question. Thus, each concept was tested equally often in each final test condition across subjects.

Two feedback messages were created for each definition question: a correct answer message and an explanation message. The correct answer message consisted of a statement of the correct answer, whereas the explanation message consisted of the correct answer as well as two additional sentences elaborating on the correct answer. The two additional sentences in the explanation feedback message were taken from the passage and helped to explain the concept. The explanation feedback did not contain any new information and it did not provide the answer to the inference question. Appendix B contains sample questions and feedback.

Procedure. The experiment consisted of two sessions spaced 2 days apart. Individual PCs running MediaLab software (Jarvis, 2004a, 2004b) were used to present all the materials and collect the responses. In Session 1, subjects were randomly assigned to one of the three feedback conditions (no feedback, correct answer feedback, and explanation feedback). Regardless of condition, they studied the 10 passages in a random order determined by the computer. Each passage was divided into two paragraphs, and each paragraph was presented for 80 s (pilot testing showed this amount

of time to be sufficient to read the entire paragraph once). Next, subjects engaged in a distractor task for 5 min (solving visuospatial puzzles). Finally, they completed a self-paced short-answer test on the critical concepts that consisted of the 20 definition questions. The questions were presented one at a time in a random order, and subjects were required to generate a response to each question. If they did not know the answer, they were instructed to make a plausible guess. Immediately after answering each question, subjects received the type of feedback that they had been assigned (no feedback, correct answer feedback, or explanation feedback). The question was always re-presented with the feedback message to provide context. Feedback was provided regardless of whether the response was correct or incorrect, and subjects were required to study the message for 20 s. In Session 2, subjects returned after 2 days to take a final short-answer test that contained 20 questions: 10 definition questions that were repeated from the initial test and 10 new inference questions. As on the initial test, questions were presented one at a time in a random order, answering was self-paced, and subjects were required to respond to each question.

Results

All results were significant at the .05 level unless otherwise stated. Pairwise comparisons were Bonferroni-corrected to the .05 level. Eta-square and Cohen's d are the measures of effect size reported for all significant effects in the analysis of variance (ANOVA) and t -test analyses, respectively.

Scoring. Two coders independently coded the responses as correct or incorrect according to a scoring rubric. Both coders were blind to condition and coded all the responses for a given question together to increase consistency. Cohen's kappa was calculated to assess interrater reliability. Reliability was high ($\kappa = .89$), and the first author (ACB) resolved the disagreements in scoring.

Initial test performance. Initial test performance was relatively low (grand $M = .43$), which was desirable for investigating the effects of feedback. A one-way ANOVA showed no effect of feedback condition ($F < 1$).

Final test performance. Figure 1 shows the proportion of correct responses on the final test as a function of feedback condition on the initial test and type of final test question. When subjects received correct answer or explanation feedback on the initial test, they performed better on the repeated definition questions relative to when they did not receive feedback. A one-way ANOVA confirmed this observation by revealing a significant main effect of type of feedback, $F(2, 57) = 6.54$, $MSE = .05$, $\eta^2 = .19$. Follow-up pairwise comparisons showed that both the correct answer and explanation feedback conditions led to a greater proportion of correct responses on repeated questions relative to the no-feedback condition (.62 vs. .43), $t(38) = 2.63$, $SED = .07$, $d = .85$; and (.66 vs. .43), $t(38) = 3.34$, $SED = .07$, $d = 1.06$, respectively. However, there was no significant difference between the correct answer and explanation feedback conditions ($t < 1$).

On the new inference questions, subjects performed best when they had received explanation feedback relative to when they got correct answer or no feedback. A one-way ANOVA showed a significant main effect of type of feedback, $F(2, 57) = 6.55$, $MSE = .04$, $\eta^2 = .19$. Pairwise comparisons confirmed that the explanation feedback condition produced a significantly greater

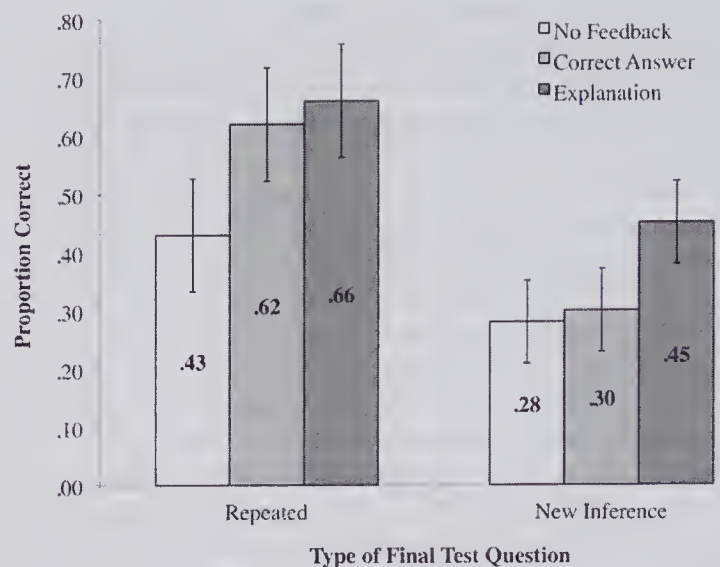


Figure 1. Proportion of correct responses on the final test as a function of feedback condition on the initial test and type of final test question in Experiment 1. Error bars represent 95% confidence intervals.

proportion of correct responses on the new inference questions relative to both the correct answer feedback and no-feedback conditions (.45 vs. .30), $t(38) = 3.13$, $SED = .05$, $d = .90$; and (.45 vs. .28), $t(38) = 2.97$, $SED = .05$, $d = 1.09$, respectively. The correct answer and no-feedback conditions did not differ ($t < 1$). In addition, an item analysis was conducted for the critical comparison between the correct answer and explanation feedback conditions by computing a t test with items as the unit of observation instead of subjects. This item analysis confirmed that explanation feedback produced superior transfer, $t(40) = 2.04$, $SED = .07$, $d = .61$.

Discussion

The results of Experiment 1 showed that the benefits of explanation feedback depend on how learning is assessed. Replicating the findings of previous studies, explanation feedback produced equivalent performance relative to correct answer feedback when retention was assessed with repeated questions on the final test (e.g., Gilman, 1969; Kulhavy et al., 1985; Mandernach, 2005; Pridemore & Klein, 1995; Sassenrath & Gaverick, 1965; Smits et al., 2008; Whyte et al., 1995). However, when the final test assessed understanding by requiring subjects to transfer their knowledge of the concept to a new context, explanation feedback led to better performance than correct answer feedback. If it can be replicated, this novel finding is important because it opens the door to a promising new direction for future research: the use of elaborative feedback to promote transfer of learning.

Experiment 2

One of the goals of Experiment 2 was to replicate the novel finding from Experiment 1 that explanation feedback produced better transfer to new inference questions than did correct answer feedback. A second goal was to investigate a potential explanation for this finding. As described in the introduction, the ability to transfer knowledge to new contexts requires understanding; however, transfer also requires retention, especially if the ability to

transfer knowledge is assessed after a delay, such as in Experiment 1. One way of conceptualizing the process of transfer involves breaking it down into three steps: (1) The learner must *recognize* that previously acquired knowledge is relevant, (2) the learner must *recall* that knowledge, and (3) the learner must *apply* that knowledge to the new context (see Barnett & Ceci, 2002). In this conceptualization, the first two steps in the transfer process reflect retention of knowledge, whereas the third step reflects understanding.

In Experiment 1, the first step (recognition) was unlikely to have been a problem: All subjects were instructed that the final test questions were about information that they had read in the passages, and therefore they recognized that they had to recall and apply their knowledge about the passages. Thus, the difference between the two feedback conditions in the ability to transfer knowledge must have been due to differences in recall, application, or both of these steps. Each explanation feedback message re-presented some information from the passage about the critical concept, so it is possible that this re-presentation boosted later recall of that information on the final test. In contrast, subjects who received correct answer feedback might have been less likely to recall this information because they had only studied it once when they read the passages. Although it is possible that differences in recall (Step 2) may have contributed to the results, we believe it is more likely that explanation feedback fostered a deeper understanding of the concepts, which facilitated the application of that knowledge to complete the final step.

In order to investigate this idea, a second phase was added to the final test in which all subjects reanswered the inference questions with the explanation feedback present (i.e., regardless of whether they had received explanation or correct answer feedback on the initial test). The rationale for the inclusion of the “reanswer” phase was that it would separate the recall and application steps in the transfer process (see Table 1 for a schematic explanation of the logic). As described above, any difference in performance between the correct answer and explanation feedback conditions when answering the new inference questions could be due to recall, application, or both of these components. By allowing subjects to consult the explanation feedback during the subsequent reanswer phase, the need to retain the information would be eliminated. Thus, any difference in performance between the two feedback conditions in the reanswer phase would reflect the subjects’ ability to apply their knowledge (i.e., their depth of understanding).

In addition to the inclusion of the reanswer phase, a few other changes were made for Experiment 2. First, the type of feedback variable was manipulated within subjects in order to show that this finding would generalize across experimental designs. Second, the no-feedback condition was dropped in order to maximize the number of items in the explanation and correct answer feedback conditions. Third, the final test consisted of only new inference questions (i.e., no repeated questions) in order to focus on replicating the key finding from Experiment 1.

Method

Participants. Twenty-four Duke University students participated for either course credit or a payment. One additional subject was excluded for not following the instructions.

Design. A single variable (type of feedback: correct answer, explanation) was manipulated within subjects, between materials.

Materials. The materials from Experiment 1 were used again.

Procedure. The procedure was the same as Experiment 1 except for the following changes. First, subjects received correct answer feedback on 10 of the definition questions on the initial test and explanation feedback for the other 10 questions. Second, the final test consisted of 20 new inference questions (no questions were repeated from the initial test). Third, the final test consisted of two phases. In Phase 1, subjects answered the new inference questions in the same manner as Experiment 1. In Phase 2, they were given the opportunity to reanswer each inference question while also viewing the relevant explanation feedback (i.e., regardless of whether they had seen the explanation feedback on the initial test or not). Subjects were told that they could re-enter their initial response or modify their response based on the information presented in the explanation feedback.

Results

Scoring. Again, two coders independently scored the responses. Reliability was almost perfect ($\kappa = .98$), and the first author (ACB) resolved the few disagreements.

Initial test performance. Overall, subjects correctly answered a little less than half the questions (grand $M = .44$), and there was no significant difference between the two feedback conditions ($t < 1$).

Final test performance. The left panel of Figure 2 shows the proportion of correct responses on the initial answer phase of the final test as a function of feedback condition on the initial test.

Table 1
The Logic Behind the Two-Phase Final Test Used in Experiment 2

Step in the transfer process	Final test phase	
	Initial answer to new inference question	Reanswer with explanation feedback
Recognition	EX = CA	EX = CA
Recall	EX > CA	EX = CA
Application	EX > CA	EX > CA

Note. When initially answering the new inference questions, there should be no difference between the two feedback conditions with respect to the recognition component of the feedback process; however, the explanation feedback condition could lead to better recall and/or application. In the reanswer phase, both recognition and recall are equated; thus, the superiority of explanation feedback over correct answer feedback must be due to the application component. EX = explanation feedback; CA = correct answer feedback.

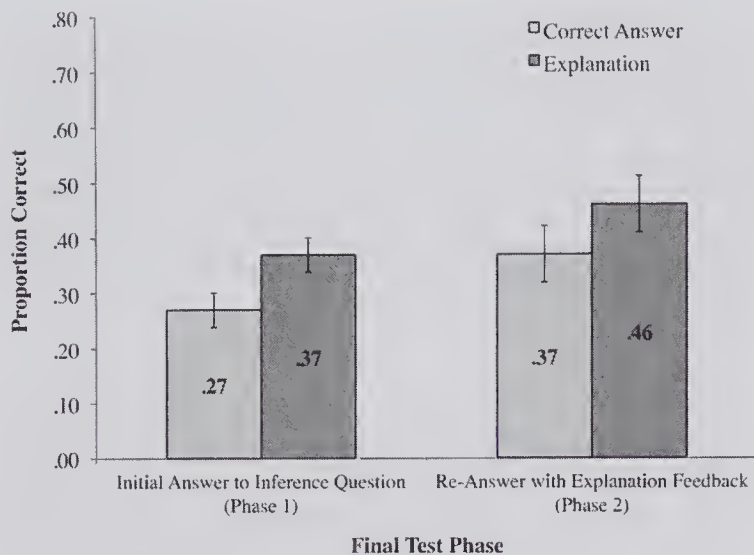


Figure 2. Proportion of correct responses on the final test as a function of feedback condition on the initial test in the initial answer (left side) and reanswer (right side) phases of the final test in Experiment 2. Error bars represent 95% confidence intervals.

Replicating the key result from Experiment 1, explanation feedback led to a significantly greater proportion of correct responses on the new inference questions relative to correct answer feedback (.37 vs. .27), $t(23) = 4.18$, $SED = .02$, $d = .50$.

The right panel of Figure 2 shows the proportion of correct responses on the reanswer phase of final test as a function of feedback condition on the initial test. Overall, the opportunity to reanswer the inference questions with the explanation feedback present improved in both the explanation feedback and correct answer feedback conditions; however, explanation feedback still produced a significantly greater proportion of correct responses than correct answer feedback (.46 vs. .37), $t(23) = 2.64$, $SED = .04$, $d = .37$. In order to compare performance on the two phases, a 2 (final test phase: initial answer, reanswer) \times 2 (type of feedback: correct answer, explanation) ANOVA was conducted. This analysis revealed significant main effects of final test phase, $F(1, 23) = 14.50$, $MSE = .02$, $\eta^2 = .39$, and type of feedback, $F(1, 23) = 15.86$, $MSE = .01$, $\eta^2 = .41$. However, the interaction was not significant ($F < 1$). In addition, an item analysis was conducted by computing the same 2 \times 2 ANOVA with items as the unit of observation instead of subjects. This item analysis revealed the same pattern of results: significant main effects of phase, $F(1, 19) = 18.45$, $MSE = .01$, $\eta^2 = .15$, and type of feedback, $F(1, 19) = 5.98$, $MSE = .01$, $\eta^2 = .15$, but no significant interaction ($F < 1$).

Discussion

Experiment 2 replicated the key novel finding from Experiment 1. When subjects received explanation feedback on the initial test, they were more successful at transferring their knowledge on the new inferences questions than when they received correct answer feedback. The additional information contained in the explanation feedback message fostered better understanding of the critical concepts, which enabled subjects to apply this knowledge to answer new inference questions. Importantly, this result also shows that the effect generalizes across experimental design—type of

feedback was manipulated between subjects in Experiment 1 and within subjects in Experiment 2.

When subjects had the opportunity to reanswer the inference questions with the explanation feedback present, the results were intriguing. Performance improved in the explanation feedback condition, which suggests that some of the information from the feedback had been forgotten; once subjects were re-presented with this information, they were able to successfully apply this knowledge to answer the inference questions. Performance also improved in the correct answer feedback condition. This improvement presumably also reflects the recall component of the transfer process—because subjects did not receive the explanation feedback on the initial test, they may not have retained this information (unless they remembered it from the passage). Practically speaking, this finding is important because it shows that giving explanation feedback after a delay can still help to improve transfer, which is consistent with recent research that shows a benefit of feedback even when its presentation is delayed (e.g., A. C. Butler, Karpicke, & Roediger, 2007; Metcalfe, Kornell, & Finn, 2009).

Most importantly, the difference in performance between the two feedback conditions for the initial answers to the new inference questions was also observed in the subsequent reanswer phase. In both feedback conditions, the presence of the explanation feedback while reanswering the inference questions meant that the burden to recall this information was removed, and any difference between the two conditions had to be due to their ability to apply their knowledge. Receiving explanation feedback on the initial test may have enabled subjects to acquire a deeper understanding of the critical concepts, which helped them to correctly answer more inference questions in the reanswer phase. Furthermore, this finding suggests that it may be particularly important to receive the explanation feedback soon after retrieving a concept from memory because the difference between the two feedback conditions persisted in the reanswer phase when the explanation feedback was always present. We turn now to discussing the importance of these findings in the context of the broader feedback literature.

General Discussion

The present research helps to resolve a paradox about elaborative feedback. Although elaborative feedback is assumed to benefit learners and it is often included in instructional methods (e.g., Corbett et al., 1997; Gibbons & Fairweather, 1998), reviewers of the feedback literature had concluded that increasing the complexity of the feedback message does not benefit learning (e.g., Bangert-Drowns et al., 1991; Kulhavy & Stock, 1989). With respect to the existing evidence in the literature, this conclusion was warranted—many studies that have isolated the effects of greater feedback complexity have found no benefit of elaborative feedback relative to correct answer feedback (e.g., Andre & Thieman, 1988; Gilman, 1969; Kulhavy et al., 1985; Peeck, 1979; Pridemore & Klein, 1995; Sassenrath & Gaverick, 1965; Whyte et al., 1995). However, all of these studies assessed retention of the correct response to a previously presented question rather than deeper understanding of the material. When understanding was assessed in the present study, explanation feedback produced better performance than correct answer feedback. This finding suggests the need for a fundamental reevaluation of how elaborative feedback affects learning.

Why did explanation feedback produce superior performance on new inference questions relative to correct answer feedback? One might expect to find an answer to this question among the various theories that have been proposed to explain how feedback affects learning. However, many of these theories do not address this question at all because they seek to describe the effects of feedback at a more complex level than that of a single task (e.g., D. L. Butler & Winne, 1995; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Such “macrolevel” theories model the influence of feedback on various student behaviors, such as self-regulation, learning strategies, and motivation, during a continuous process of learning that includes repeated presentations of feedback. Although other theories provide a “microlevel” account of learning from feedback during a single task, these theories are either too general (e.g., Bangert-Drowns et al., 1991) or focus on explaining other feedback phenomena (e.g., the relationship between response confidence and feedback processing; Kulhavy, 1977). Kulhavy and Stock (1989) put forth the only theoretical framework that specifically addresses the effects of elaborating the feedback message beyond providing the correct answer. Despite their efforts to develop a coherent account of how elaboration affects learning, they were “unable to reach any useful conclusion regarding how the elaborative component of the feedback operates” (Kulhavy & Stock, 1989, p. 289). Recent microlevel reviews of the feedback literature describe many of these theories but offer no new ideas regarding elaborative feedback (e.g., Mory, 2004; Shute, 2008).

Given the dearth of existing feedback theory upon which to draw, we looked to theories in other domains in order to develop an explanation for our findings. One relevant theory is the framework proposed by Barnett and Ceci (2002) to explain the process of transfer and the factors that influence whether it will occur. As described above, they conceptualize the process of transfer in terms of three steps: recognition, recall, and application. Both correct answer and explanation feedback can improve the retention of specific knowledge, which would facilitate later recall of the information (i.e., the second step in the transfer process); this conclusion is supported by the finding that the two types of feedback produced equivalent performance on the definition questions that were repeated on the final test in Experiment 1. However, explanation feedback may also enable learners to better comprehend the concepts, thus facilitating the application of that knowledge to new contexts (i.e., the third step in the transfer process). The results of the reanswer phase in Experiment 2 support this conclusion. When subjects reanswered the inference questions with the explanation feedback present, the superiority of explanation feedback persisted even though the recall demands were removed, suggesting that the locus of the effect is the application step of the transfer process.

Another way of framing our findings is through the lens of text-processing theories that conceptualize the development of understanding as a process that requires representing a text on multiple levels (for a review, see Graesser, Millis, & Zwaan, 1997; Kintsch, 1998). Such theories often differentiate between three levels of representation: the *surface level*—the specific words and syntax used in the text; the *textbase*—an abstract representation of the ideas and their connections; and the *situation model*—a personal interpretation of the text that often includes preexisting knowledge. According to most theories, the situation model is the representational level that reflects deep understanding and sup-

ports the transfer of knowledge. Within the context of the present study, processing the explanation feedback after an initial retrieval attempt may have helped subjects to improve their situation model of the text and achieve a deeper understanding. A more developed situation model would be expected to enable superior transfer of knowledge to the new inference questions, which were aligned with this representational level. In contrast, the repeated questions used to assess retention were aligned with memory for the textbase, and thus explanation feedback would not be expected to benefit performance on these items relative to correct answer feedback.

One remaining puzzle is why explanation feedback was effective at facilitating understanding when it was given on the initial test, but it did not have the same effect on the correct answer condition when it was presented during the reanswer phase of the final test. Although additional research will be needed to further explore this finding, one potential explanation revolves around the concept of memory reconsolidation. In general, practice retrieving the critical concepts from memory would be expected to help subjects to better retain these concepts and transfer them to new contexts, regardless of feedback condition (e.g., A. C. Butler, 2010; Roediger & Butler, 2011). However, retrieval may also reopen a memory so that it must be reconsolidated, meaning that the memory enters a labile state in which it can be altered (e.g., Hupbach, Gomez, Hardt, & Nadel, 2007; for a review, see Dudai, 2006; Lee, 2009). For example, a recent study by Finn and Roediger (2011) showed that postretrieval processing of new information results in the integration of this information into the existing memory, thereby enhancing retention. In the present study, postretrieval processing of the explanation feedback on the initial test may have resulted in the information being integrated into the memory of the concept, thus building a deeper understanding (i.e., a more developed situation model). Retrieval during the final test should also involve reopening the memory, giving a chance for both groups of subjects to integrate the explanation feedback presented in the reanswer phase into their memories; however, it may be that the memory must be successfully reconsolidated (over time) before a deeper understanding is developed. Although admittedly somewhat speculative, this reconsolidation hypothesis provides a potential starting point for follow-up studies.

The present findings open the door for new research that investigates the role of feedback in promoting transfer of knowledge. The need for this research is apparent with respect to all types of elaborative feedback, but also more generally with other factors that influence the efficacy of feedback. The vast majority of feedback studies in the literature use final tests with repeated questions to assess retention of knowledge. Although retention is certainly an important learning outcome, so too is understanding. Thus, there is a great need for research on how feedback affects transfer for both theoretical and pedagogical purposes. If understanding is ignored as a learning outcome, many promising methods of providing feedback may be misconceived and overlooked. For example, one method that may help to produce substantial understanding is to give students correct answer feedback and then have them generate their own explanations for why their response is correct or incorrect. Previous studies have not found a benefit of such a procedure relative to simply providing correct answer feedback (e.g., McDaniel & Fisher, 1991); however, these studies have measured retention rather than understanding. In summary, the findings of the present study indicate that transfer of knowl-

edge represents a fruitful new frontier for feedback research—it is time for feedback researchers to move beyond measuring retention and investigate how feedback affects understanding.

References

- Andre, T., & Thieman, A. (1988). Level of adjunct question, type of feedback, and learning concepts by reading. *Contemporary Educational Psychology, 13*, 296–307. doi:10.1016/0361-476X(88)90028-8
- Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213–238.
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5*, 7–74.
- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118–1133. doi:10.1037/a0019902
- Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*, 273–281. doi:10.1037/1076-898X.13.4.273
- Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a meta-cognitive error: Feedback enhances retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 918–928. doi:10.1037/0278-7393.34.4.918
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Psychology, 65*, 245–281.
- Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd ed., pp. 849–874). New York, NY: Elsevier.
- Dudai, Y. (2006). Reconsolidation: The advantage of being refocused. *Current Opinion in Neurobiology, 16*, 174–178. doi:10.1016/j.conb.2006.03.010
- Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory, 18*, 335–350. doi:10.1080/09658211003652491
- Finn, B., & Roediger, H. L., III. (2011). Enhancing retention through reconsolidation: Negative emotional arousal following retrieval enhances later memory. *Psychological Science, 22*, 781–786. doi:10.1177/0956797611407932
- Gibbons, A. S., & Fairweather, P. G. (1998). *Computer-based instruction: Design and development*. Englewood Cliffs, NJ: Educational Technology.
- Gilman, D. A. (1969). Comparison of several feedback methods for correcting errors by computer-assisted instruction. *Journal of Educational Psychology, 60*, 503–508. doi:10.1037/h0028501
- Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). Auto-Tutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*, 612–618. doi:10.1109/TE.2005.856149
- Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163–189. doi:10.1146/annurev.psych.48.1.163
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487
- Hupbach, A., Gomez, R., Hardt, O., & Nadel, L. (2007). Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. *Learning & Memory, 14*, 47–53. doi:10.1101/lm.365707
- Jarvis, B. G. (2004a). Medialab (Version 2004.2.87) [Computer software]. New York, NY: Empirisoft Corporation.
- Jarvis, B. G. (2004b). DirectRT (Version 2004.1.0.55) [Computer software]. New York, NY: Empirisoft Corporation.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. doi:10.1037/0033-2909.119.2.254
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232.
- Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279–308. doi:10.1007/BF01320096
- Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology, 10*, 285–291. doi:10.1016/0361-476X(85)90025-6
- Lee, J. L. C. (2009). Reconsolidation: Maintaining memory relevance. *Trends in Neurosciences, 32*, 413–420. doi:10.1016/j.tins.2009.05.002
- Mandernach, B. J. (2005). Relative effectiveness of computer-based and human feedback for enhancing student learning. *The Journal of Educators Online, 2*, 1–17.
- McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*, 192–201. doi:10.1016/0361-476X(91)90037-L
- McGeoch, J. A. (1942). *The psychology of human learning: An introduction*. New York, NY: Longmans, Green and Co. doi:10.2307/2262568
- Metcalf, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adult's vocabulary learning. *Memory & Cognition, 37*, 1077–1087. doi:10.3758/MC.37.8.1077
- Mory, E. H. (2004). Feedback research review. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Mahwah, NJ: Erlbaum.
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3–8. doi:10.1037/0278-7393.31.1.3
- Peeck, J. (1979). Effects of differential feedback on the answering of two types of questions by fifth- and sixth-graders. *British Journal of Educational Psychology, 49*, 87–92. doi:10.1111/j.2044-8279.1979.tb02401.x
- Phye, G. D., & Sanders, C. E. (1994). Advice and feedback: Elements of practice for problem solving. *Contemporary Educational Psychology, 19*, 286–301. doi:10.1006/ceps.1994.1022
- Plowman, L., & Stroud, J. B. (1942). Effect of informing pupils of the correctness of their responses to objective test questions. *Journal of Educational Research, 36*, 16–20.
- Pridmore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-assisted instruction. *Contemporary Educational Psychology, 20*, 444–450. doi:10.1006/ceps.1995.1030
- Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27. doi:10.1016/j.tics.2010.09.003
- Roper, W. J. (1977). Feedback in computer assisted instruction. *Programmed Learning and Educational Technology, 14*, 43–49.
- Sassenrath, J. M., & Gaverick, C. M. (1965). Effects of differential feedback from examinations on retention and transfer. *Journal of Educational Psychology, 56*, 259–263. doi:10.1037/h0022474
- Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. doi:10.3102/0034654307313795
- Smits, M. H. S. B., Boon, J., Sluijsmans, D. M. A., & van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments, 16*, 183–193. doi:10.1080/10494820701365952

Travers, R. M. W., Van Wageningen, R. K., Haygood, D. H., & McCormick, M. (1964). Learning as a consequence of the learner's task involvement under different conditions of feedback. *Journal of Educational Psychology*, 55, 167–173. doi:10.1037/h0048319

Whyte, M. M., Karolick, D. M., Neilsen, M. C., Elder, G. D., & Hawley, W. T. (1995). Cognitive styles and feedback in computer-assisted instruction. *Journal of Educational Computing Research*, 12, 195–203. doi:10.2190/M2AV-GEHE-CM9G-J9P7

Appendix A

Sample Passages Used in Experiments 1 and 2

These passages are associated with the sample questions and feedback provided in Table 1.

The Respiratory System

Humans breathe in and out anywhere from 15 to 25 times per minute. The main function of the respiratory system is gas exchange between the external environment and the circulatory system. A gas that the body needs to get rid of, carbon dioxide, is exchanged for a gas that the body can use, oxygen. The lungs are the most critical component of the respiratory system because they are responsible for the oxygenation of the blood and the concomitant removal of carbon dioxide from the circulatory system. Gas exchange occurs in tiny, thin-walled air sacs called alveoli, which lie at the end of the many branches of tubes in the lungs. Within the alveoli, gas exchange occurs as a result of diffusion. Diffusion is the movement of particles from a region of high concentration to a region of low concentration. The oxygen concentration is high in the alveoli, so oxygen diffuses across the alveolar membrane into the pulmonary capillaries, which are small blood vessels that surround each alveolus. The hemoglobin in the red blood cells passing through the pulmonary capillaries has carbon dioxide bound to it and very little oxygen. The oxygen binds to hemoglobin and the carbon dioxide is released. Since the concentration of carbon dioxide is high in the pulmonary capillaries relative to the alveolus, carbon dioxide diffuses across the alveolar membrane in the opposite direction. The exchange of gases across the alveolar membrane occurs rapidly—usually in fractions of a second.

Humans do not have to think about breathing because the body's autonomic nervous system controls it. The respiratory centers that control the rate of breathing are located in the pons and medulla oblongata, which are both part of the brainstem. The neurons that live within these centers automatically send signals to the diaphragm and intercostal muscles to contract and relax at regular intervals. Neurons in the cerebral cortex can also voluntarily influence the activity of the respiratory centers. A region within the cerebral cortex, called motor cortex, controls all voluntary motor functions, including telling the respiratory center to speed up, slow down, or even stop. However, the influence of the nerve centers that control voluntary movements can be overridden by the autonomic nervous system. Several factors can trigger such an override. One of these factors is the concentration of oxygen in the

blood. Specialized nerve cells within the aorta and carotid arteries called peripheral chemoreceptors monitor the oxygen concentration of the blood. If the oxygen concentration decreases, the chemoreceptors signal to the respiratory centers in the brain to increase the rate and depth of breathing. These peripheral chemoreceptors also monitor the carbon dioxide concentration in the blood. Another factor is chemical irritants. Nerve cells in the airways can sense the presence of unwanted substances like pollen, dust, water, or cigarette smoke. If chemical irritants are detected, these cells signal the respiratory centers to contract the respiratory muscles, and the coughing that results expels the irritant from the lungs.

Vaccines

A vaccine is a biological preparation that establishes or improves immunity to a particular disease. Most vaccines are prophylactic, which means that they prevent or ameliorate the effects of a future infection by any natural pathogen. However, vaccines have also been used for therapeutic purposes, such as for alleviating the suffering of people already afflicted with a disease. The early vaccines were inspired by the concept of variolation, which originated in Asia during the 13th century. Variolation is a technique in which a person is deliberately infected with a weak form of a disease by inhaling it through the nose or mouth. Upon recovery, the individual was immune to the disease. A small proportion of the people who were variolated died, but nowhere near the proportion that died when they contracted the disease naturally. By the 18th century, knowledge of variolation had spread to Europe where medical researchers Edward Jenner and Louis Pasteur transformed the ancient technique into the modern day practice of inoculation with vaccines. Inoculation represented a major breakthrough because it reduced the risk of vaccination, while maintaining its effectiveness. Inoculation is the practice of deliberate infection through a skin wound. This new technique produces a smaller, more localized infection relative to variolation in which inhalation of viral particles spreads the infection more widely. The smaller infection works better because it is sufficient to stimulate immunity to the virus, but it keeps the virus from replicating enough to reach levels of infection likely to kill a patient.

(Appendices continue)

Vaccines work because they prepare the immune system to deal with pathogens that it may encounter in the future. When a vaccine is given, the immune system recognizes the vaccine agents as foreign, destroys them, and then “remembers” them. When the real version of virus comes along, the body recognizes it and destroys the infected cells before they multiply. Of course, vaccines do not guarantee complete protection against the disease because sometimes a person’s immune system does not respond for various reasons. Still, even when a vaccinated individual does develop the disease vaccinated against, the disease is likely to be milder than without vaccination. Overall, the invention of vaccines has led to a marked decrease in the prevalence of deadly diseases, such as smallpox, polio, measles,

and typhoid. As long as the vast majority of people are vaccinated, it is much more difficult for an outbreak of disease to occur and spread because of herd immunity. Herd immunity describes a type of immunity that occurs when the vaccination of a portion of the population (or herd) provides protection to unvaccinated individuals. Herd immunity theory proposes that for diseases passed from person-to-person, it is more difficult to maintain a chain of infection when large numbers of a population are immune. The higher the proportion of individuals who are immune, the lower the likelihood that a susceptible person will come into contact with an infected individual. Despite potential protection from herd immunity, mainstream medical opinion is that everyone should be vaccinated.

Appendix B

Sample Questions and Feedback Taken From Passages on the Respiratory System and Vaccines, Respectively

Retention questions were used on the initial test and repeated on the final test, whereas transfer questions were only used on the final test.

The Respiratory System

Retention Question: What is the process by which gas exchange occurs in the part of the human respiratory system called the alveoli?

Correct Answer Feedback: *Gas exchange occurs within the alveoli through diffusion.*

Explanation Feedback: *Gas exchange occurs within the alveoli through diffusion. Diffusion is the movement of particles from a region of high concentration to a region of low concentration. The oxygen concentration is high in the alveoli and the carbon dioxide concentration is high in the pulmonary capillaries, so the two gases diffuse across the alveolar membrane in opposite directions towards lower concentrations.*

Inference Question: If people are having trouble breathing, they are often given pure oxygen to inhale. How does breathing pure oxygen facilitate gas exchange relative to regular air?

Answer: *Breathing pure oxygen increases the oxygen concentration in the alveoli, so oxygen will diffuse more rapidly across the alveolar membrane into blood in the pulmonary capillaries.*

Vaccines

Retention Question: What vaccination technique did Edward Jenner and Louis Pasteur develop that improved upon the ancient practice of variolation?

Correct Answer Feedback: *Edward Jenner and Louis Pasteur developed the technique of inoculation to improve upon the ancient practice of variolation.*

Explanation Feedback: *Edward Jenner and Louis Pasteur developed the technique of inoculation to improve upon the ancient practice of variolation. Inoculation is the practice of deliberate infection through a skin wound, whereas variolation involves inhaling a weak form of the disease. The new technique produces a smaller, more localized infection that is adequate to stimulate immunity to the virus, but keeps it from replicating enough to be dangerous.*

Inference Question: The recently developed nasal spray flu vaccine, which is inhaled through the nose, contains weakened viruses that only cause infection at the cooler temperatures found within the nose. In what sense does this new method of vaccination combine the techniques of inoculation and variolation?

Answer: *The nasal spray flu vaccine is similar to inoculation in that it produces a smaller, more localized infection, but also like variolation in that the virus is inhaled.*

Received May 16, 2011

Revision received September 24, 2012

Accepted October 11, 2012 ■

Note-Taking With Computers: Exploring Alternative Strategies for Improved Recall

Dung C. Bui, Joel Myerson, and Sandra Hale
Washington University in St. Louis

Three experiments examined note-taking strategies and their relation to recall. In Experiment 1, participants were instructed either to take organized lecture notes or to try and transcribe the lecture, and they either took their notes by hand or typed them into a computer. Those instructed to transcribe the lecture using a computer showed the best recall on immediate tests, and the subsequent experiments focused on note-taking using computers. Experiment 2 showed that taking organized notes produced the best recall on delayed tests. In Experiment 3, however, when participants were given the opportunity to study their notes, those who had tried to transcribe the lecture showed better recall on delayed tests than those who had taken organized notes. Correlational analyses of data from all 3 experiments revealed that for those who took organized notes, working memory predicted note-quantity, which predicted recall on both immediate and delayed tests. For those who tried to transcribe the lecture, in contrast, only note-quantity was a consistent predictor of recall. These results suggest that individuals who have poor working memory (an ability traditionally thought to be important for note-taking) can still take effective notes if they use a note-taking strategy (transcribing using a computer) that can help level the playing field for students of diverse cognitive abilities.

Keywords: note-taking, note quantity and quality, computers, individual differences, working memory

Note-taking has long been linked to positive test performance (e.g., Armbruster, 2000; Crawford, 1925b), and this relationship is not lost on students, who acknowledge that lecture note-taking is a crucial component of the educational experience (Dunkel & Davy, 1989). In fact, lecturing constitutes nearly 83% of college instructors' teaching methods (Wirt et al., 2001), and nearly all college students take notes in class (Palmatier & Bennett, 1974), even when they are not explicitly told to do so by the instructor (Williams & Eggert, 2002). Researchers have identified two primary ways in which classroom note-taking is beneficial: *Encoding* and *external storage* (Di Vesta & Gray, 1972). The encoding benefit (also termed the process benefit) refers to the learning that results from the act of taking notes, whereas the external storage benefit (also termed the product benefit) refers to the benefit that comes from studying the notes. Furthermore, Kiewra (1985) pointed out that utilizing *both* aspects of note-taking in conjunction provides a more potent learning tool than either aspect on its own (e.g., Fisher & Harris, 1973; Kiewra, DuBois, Christensen, Kim, & Lindberg, 1989).

Recent advancements in technology have led to more computers being introduced into the classroom and incorporated into students' learning experiences, and the availability of portable computers has resulted in a steady increase in the percentage of college students who own one (89%; Smith & Caruso, 2010). Research has

compared typing speed to writing speed and found evidence that proficient typists can type faster than they can handwrite (e.g., Brown, 1988), and that this pattern emerges in children as young as sixth grade (Rogers & Case-Smith, 2002). Thus, it would appear that for many students, portable computers can increase their transcription speed when they take lecture notes.

The Relation Between Working Memory and Note-Taking

Despite its benefits, lecture note-taking is a complex and cognitively demanding skill that requires comprehending what the instructor is saying, holding that information in memory, organizing and paraphrasing it, and then writing it down before it is forgotten, all while attending to the ongoing lecture. When note-taking skill is framed as a composition of more basic cognitive abilities, it is clear that one reason why students' notes vary among one another is likely because of individual differences in these lower-order abilities.

One ability hypothesized to be important in note-taking is *working memory* (e.g., Olive & Piolat, 2002), the ability to temporarily hold and manipulate a limited amount of information (Baddeley, 1986). While some studies report a correlation between working memory and note-taking (e.g., Kiewra & Benton, 1988; Kiewra, Benton, & Lewis, 1987), other studies do not (e.g., Cohn, Cohn, & Bradley, 1995; Peverly et al., 2007). It is possible that these mixed results are due to variability in the note-taking strategies that students naturally use. Without explicit instructions, students may choose strategies that vary in the extent to which they rely on working memory, potentially masking a correlation between working memory and note-taking.

This article was published Online First October 8, 2012.

Dung C. Bui, Joel Myerson, and Sandra Hale, Department of Psychology, Washington University in St. Louis.

Correspondence concerning this article should be addressed to Dung C. Bui, Department of Psychology, Washington University in St. Louis, St. Louis, MO 63130. E-mail: dcui@wustl.edu

Currently, it is unclear whether working memory always plays a vital role in note-taking, or whether working memory is important only for select note-taking strategies. Nonetheless, if note-taking, like other cognitive skills, relies on basic processing abilities, then it would not be too surprising if individual differences in such abilities account for much of the variance in note-taking as it relates to test performance. And if taking notes in and of itself provides an encoding benefit, then individual differences in working memory may predict test performance even when individuals do not get to study their notes.

The Relation Between Note-Quantity and Recall

As mentioned previously, note-taking is beneficial in and of itself, independent of studying. One consequence of this is that taking more notes may lead to better learning, as more information has been encoded. Indeed, studies have shown a significant relation between note-quantity and test performance, both when students study their notes (e.g., Crawford, 1925a; Kiewra & Benton, 1988), as well as when they are not allowed to study their notes (e.g., Fisher & Harris, 1973). Moreover, Peverly et al. (2007) found that measures of transcription fluency (how fast one can take notes) predicted note-taking, which in turn predicted performance on writing tasks measuring recall, raising the possibility that increasing transcription fluency may be one way to increase recall.

The benefit of simply writing down what is said in lecture may be explained by the *generation effect* (e.g., Rabinowitz & Craik, 1986): the finding that information is better remembered when it is generated, compared to when it is simply read or heard. One interpretation of this effect is that the act of generation is itself really an act of recall, and that the opportunity for recall benefits memory for that information (Slamecka & Graf, 1978). Similarly, Scardamalia and Bereiter (1986) suggested that note-taking forces students to generate (recall) information just heard during the lecture, which benefits memory more than merely hearing the information. Faber, Morris, and Lieberman (2000) found that note-taking without an opportunity to study the notes facilitated comprehension for students as young as ninth grade, providing further support for the role of generation in lecture note-taking (see Kobayashi, 2005, for a meta-analysis focusing on the factors that influence the encoding benefits of note-taking).

Conway and Gathercole's (1990) *translation hypothesis* provides another account of why writing down what is said in lecture can benefit memory. According to their hypothesis, when input activities require translation between specialized processing domains, this leads to the formation of more distinctive memory representations. Because listening to a lecture requires phonological processing, whereas writing down what was said invokes orthographical processing, the translation effect should benefit memory. Moreover, the translation hypothesis provides an intriguing explanation as to why quantity of notes is positively correlated with test performance: Writing down more of what was said in a lecture leads to more information being encoded, as well as more distinctive memory representations.

Alternatively, the quality of one's notes may be much more important than the quantity. This view is consistent with the *levels of processing framework* in that taking organized notes would appear to involve the kind of semantic processing that leads to better retention of verbal information (Craik & Lockhart, 1972).

Taking organized notes may also enhance retention because it involves the kind of "desirable difficulty" highlighted by Bjork (1994). Although there is an extensive literature that examines the contributions of the quantity and quality of students' notes to learning and test performance, perhaps not surprisingly, previous studies have been focused almost exclusively on taking handwritten notes (for a meta-analytic review, see Kobayashi, 2005). It is possible, however, that taking notes with a computer may change the balance between note quality and quantity, and the current study is the first to directly examine this emerging issue.

The Current Study

If transcription speed plays an important role in note-taking (e.g., Peverly et al., 2007) and typing is faster than writing by hand (e.g., Brown, 1988), then computers would appear to provide an opportunity to increase note-quantity for virtually all students, thereby improving their test performance. Moreover, given that note-quantity predicts test performance, it is possible that instructing students to take as much notes as possible will prove more beneficial for learning than the usual practice of taking organized notes.

This study has three aims. One aim is to compare taking notes by hand with taking notes using a computer in terms of their effects on test performance. The second aim is to compare the effects of taking organized notes with the effects of trying to transcribe a lecture. The third aim is to examine the role of working memory in these two note-taking strategies. Experiment 1 compares the effects of different note-taking methods (hand or computer) and note-taking instructions (organizing or transcribing) on an immediate test. Experiments 2 and 3 examine how these same two note-taking strategies influence performance on delayed tests when people are not allowed to study their notes (Experiment 2), and when they are given an opportunity to study (Experiment 3). Finally, we examine the role of individual differences in working memory in determining who benefits from different note-taking strategies.

Experiment 1

Method

Participants. Eighty undergraduate students (53 females and 27 males; mean age = 19.2 years, $SD = 1.2$), all proficient English speakers, participated for course credit.

Materials. Participants were tested individually in a private testing room equipped with a PC and a 15-in. (38.1-cm) monitor that was used for stimulus presentation on all tasks. Note-taking was done using either pen and paper or computer and keyboard, depending on the condition. On the free recall and short answer tests, all participants responded using the computer keyboard.

Reading span task. A reading span task (Daneman & Carpenter, 1980) was used to assess working memory ability. Participants were shown a series of sentences and digits. After reading each sentence aloud, participants reported whether or not the sentence was sensible, at which time the sentence disappeared and a digit appeared on the screen, and participants read the digit aloud. At the end of each series, participants were cued to recall the digits aloud

in the order of presentation. The total score was calculated by summing the series lengths of the correct trials.

Lexical decision task. Processing speed was measured using a lexical decision task in which participants were shown strings of letters (e.g., “bin,” “mun”). For each letter string, participants made a decision as to whether or not it was a real English word. Each individual’s measure of processing speed was based on correct responses to both words and nonwords.

Lecture. Participants listened to an 11-min lecture that consisted of a passage from a nonfiction book (Carnes, 1995) in which a popular film from the 1930s (*The Charge of the Light Brigade*) is compared with the event it depicted (the Crimean War). None of the participants in any of the three experiments in the present study had ever seen the film, and they did not know anything about the Crimean War. This passage was used previously by Rawson and Kintsch (2005), who developed a scoring system in which select idea units represented main points, important details, or unimportant details of the passage. Of the 125 total idea units, eight were classified as representing main points, 15 represented important details, and 16 represented unimportant details (Rawson & Kintsch, 2005). The 1,541-word lecture was read aloud and recorded in a sound-proof room at an average rate of 140 words per minute. The recording was subsequently presented to participants through the computer speakers.

Tests. Two types of test, free recall and short answer, were used to assess memory for the passage. The short answer test (Rawson & Kintsch, 2005) consisted of 18 questions (e.g., “What was the political idea that *Light Brigade* was intended to promote?”), of which eight were about important details, and 10 were about unimportant details.

Design and procedure. A 2 (instruction: organize, transcribe) \times 2 (method: hand, computer) between-subjects design was used. For this and all subsequent experiments, analyses of variance (ANOVAs) were performed on the notes, free recall performance, and short answer performance, and planned comparisons with Bonferroni corrections were conducted for all significant interactions.

Following collection of demographic information, participants performed the tasks in the following order: Reading span task, lexical decision task, lecture note-taking, free recall test, and short answer test. Participants were told that they would be listening to the lecture and were instructed to take notes for an upcoming test. Further instructions were given as to how the notes should be taken. For those in the *organize* condition, participants were told to paraphrase and to organize their notes as much as possible. Those

in the *transcribe* condition were told to record as much of the lecture as possible. Finally, participants in the *hand* condition were provided a notepad and a pen, and participants in the *computer* condition were told to type their notes into a computer file using a word processor.

When participants finished listening to the lecture, the experimenter made the notes unavailable to the participants and administered the two tests. For the free recall test, participants were told that they had 10 min to recall as much information as they could remember from the lecture. This was followed by the short answer test, which participants also had 10 min to complete. Two independent raters, blind with respect to the conditions, scored all of the notes and free recall responses. Participants were given either a full point for recall of an entire idea unit, half of a point for partial recall of the idea unit, or zero for no recall. Inter-rater reliability was .85 and .82 for notes and free recall responses, respectively. Discrepancies in scoring were resolved by taking the average of the scores given by the raters.

Results

The groups assigned to the four conditions did not differ in either working memory or processing speed, both $F_s < 1.63$, precluding the possibility that any group differences in note-taking and test performance could be due to differences in these cognitive abilities.

Note-taking. For each participant, note-quantity was measured as the proportion of the idea units from the lecture that were recorded in the participant’s notes (see Table 1). There was an effect of method on note-quantity, $F(1, 76) = 17.68, p < .001, \eta^2 = .19$, as well as an effect of instruction, $F(1, 76) = 4.07, p < .05, \eta^2 = .05$, indicating that, on average, notes taken with a computer contained a larger proportion of idea units than handwritten notes and transcribed notes contained a larger proportion of idea units than organized notes. There was also an interaction between note-taking method and instruction, $F(1, 76) = 4.07, p < .05, \eta^2 = .04$, reflecting the fact that when using a computer, the instruction to try and transcribe the lecture was associated with a larger proportion of idea units than the instruction to try and take organized notes, $t(38) = 2.71, p < .05$, whereas there was no effect of note-taking instructions on the proportion of idea units when notes were taken by hand, $t < 1.00$.

Free recall. Table 1 displays the overall proportion of idea units, as well as the proportions of main points, important details, and unimportant details recalled by each group. Those who took

Table 1
Experiment 1: Proportions of Idea Units Recalled (Standard Deviations in Parentheses)

Group	Note-taking overall	Free recall				Short answer		
		Overall	Main ideas	Important details	Unimportant details	Overall	Important details	Unimportant details
Hand								
Organize	.28 (.12)	.12 (.05)	.17 (.10)	.18 (.09)	.10 (.08)	.47 (.19)	.52 (.16)	.42 (.26)
Transcribe	.28 (.10)	.12 (.03)	.17 (.12)	.21 (.10)	.08 (.07)	.46 (.15)	.45 (.17)	.47 (.18)
Computer								
Organize	.34 (.13)	.12 (.05)	.21 (.14)	.16 (.10)	.10 (.10)	.50 (.20)	.53 (.20)	.46 (.25)
Transcribe	.44 (.12)	.18 (.06)	.25 (.13)	.24 (.12)	.12 (.08)	.64 (.12)	.72 (.16)	.58 (.13)

notes using a computer recalled more idea units than those who took handwritten notes, $F(1, 76) = 7.62, p < .01, \eta^2 = .08$, and those who took transcribed notes recalled more than those who took organized notes, $F(1, 76) = 7.82, p < .01, \eta^2 = .08$. There was an interaction between method and instruction, $F(1, 76) = 6.41, p < .05, \eta^2 = .06$: Transcribing led to better overall free recall performance than taking organized notes when notes were taken using a computer, $t(38) = 3.36, p < .05$, whereas there was no effect of instruction when notes were taken by hand, $t < 1.00$. Analysis of free recall of main idea units indicated an effect of method: Taking notes using a computer led to greater recall of main idea units compared to taking notes by hand, $F(1, 76) = 4.73, p < .05, \eta^2 = .06$. However, there was no effect of instruction and no interaction between the two factors, $F_s < 1.00$. Recall of important details was better for those who transcribed, $F(1, 76) = 5.61, p < .05$, but did not differ by method, and there was not an instruction by method interaction, $F_s < 1.00$. Finally there were no effects of method or instruction on unimportant details and no interaction, $F_s < 1.00$.

Short answer. Table 1 shows overall performance, as well as performance on short answer questions addressing important and unimportant details. Taking notes using a computer led to better overall test performance compared to taking notes by hand, $F(1, 76) = 7.69, p < .01$, and taking transcribed notes led to better performance compared to taking organized notes, $F(1, 76) = 3.46, p < .05$. An interaction between method and instruction was found, $F(1, 76) = 3.97, p < .05$, reflecting a difference in performance between the two note-taking instructions when notes were taken with a computer, $t(38) = 2.76, p < .01$, but not when notes were taken by hand, $t < 1.00$. Analysis of recall of important details revealed an effect of method, $F(1, 76) = 12.85, p < .01, \eta^2 = .13$, indicating that using computers led to better performance than taking notes by hand, but no effect of instruction, $F < 2.00$. A significant interaction between method and instruction was found, $F(1, 76) = 10.64, p < .01, \eta^2 = .10$, reflecting the fact that when taking notes by computer, transcribing led to better performance than taking organized notes on questions addressing important details, $t(38) = 3.15, p < .01$, whereas when taking notes by hand, there was no difference between the two note-taking strategies, $t < 1.50$. With respect to recall of unimportant details, there were no significant main effects and no interaction between the two factors, $F_s < 3.0$.

Discussion

When people used a computer to take notes, they took more notes and recalled more of the lecture than when they took notes by hand. Moreover, when they used a computer and were instructed to try and transcribe the lecture, this strategy was associated with the most notes and the best performance on both the free recall and short answer tests, with performance not only exceeding that of those who took organized notes with a computer but also that of those who used either handwritten note strategy. And because the benefits of transcribing with a computer extended to recall of both the main idea units and the important details, it is clear that the superior overall performance of those using this strategy was not simply due to their including more unimportant information in their notes or in their free recall. The present results are consistent with the generation effect (Slamecka & Graf, 1978)

as well as the translation hypothesis (Conway & Gathercole, 1990), both of which would predict that memory should be better for information that is written down compared to information that is simply heard.

Interestingly, for people taking notes by hand, telling them to write down as much as possible from the lecture did not result in more notes compared to telling them to paraphrase and to organize their notes. One possible explanation is that this is simply because of the physical limitations imposed by handwriting. In other words, it is possible that an individual transcribing notes by hand cannot physically write fast enough, or for a long enough period of time, to produce more notes than someone who is organizing by hand. This highlights the potential impact that computers can have on note-taking in classroom settings, as keyboards allow for faster note-taking for a longer period of time.

The ability to take more notes, of course, provides clear benefits for students from an external storage standpoint because it means there is more information to study. However, participants in Experiment 1 were not allowed to study their notes, and thus differences in external storage cannot explain any of the observed differences in test performance. Instead, it would seem more likely that the differences between groups were driven by the encoding benefit that comes from note-taking. Our results for both the free recall and short answer tests are consistent with what would be predicted based on encoding benefits—transcribing with a computer led to more notes and thus to superior memory performance. Taken together, the results of Experiment 1 indicate that transcribing lecture notes using a computer not only yields a greater quantity of notes, but also results in a benefit on both free recall and short answer tests.

One potential concern about recommending to students, based on this finding, that they try transcribing lectures (rather than taking organized notes) is that students might do so at the expense of failing to highlight important details (i.e., note quality could suffer). Our results suggest that this does not seem to be the case, at least for those who took notes using a computer: Indeed, the fact that the transcription strategy, when combined with using a computer, resulted in not only the most notes but also recall of the most main ideas and important details of any of the four groups in the experiment indicates that the resulting greater quantity of information did not come at the expense of the quality of the information.

Nevertheless, people may process the information more deeply when they organize their notes. If it is the case that there are differences in the level of processing produced by the two note-taking instructions, then clear-cut predictions can be made about long-term retention of the lecture material. Specifically, it would be expected that any advantage of the transcription strategy over taking organized notes would change over time, such that taking organized notes would lead to better long-term learning. This would be consistent with the levels-of-processing framework, which predicts that deeper encoding of information will lead to better long-term retention than shallow encoding (Craik & Lockhart, 1972).

Experiment 2

Of primary interest in Experiment 1 was the finding that taking transcribed notes using a computer led to better immediate test

performance than taking organized notes. As just noted, however, transcription likely involves shallower processing than taking organized notes, and thus according to the levels-of-processing framework, the results of Experiment 1 should not generalize to situations in which the test is delayed. Accordingly, Experiment 2 was designed to test the predictions of the levels-of-processing framework by examining how note-taking instructions affect performance on both an immediate test (a partial replication of Experiment 1) and a test administered 24 hr after the lecture. If taking organized notes is associated with deeper processing, then a significant delay by instruction interaction would be expected, reflecting greater forgetting by those who were instructed to try and transcribe the lecture. Because effects of note-taking instruction were only found for those who used computers, and because computer use led to the best performance overall, all of the participants in Experiment 2 took notes using a computer.

Method

Participants. Seventy-six undergraduate students (37 females and 39 males; mean age = 19.4 years, $SD = 1.3$), all of whom were proficient English speakers, participated for course credit.

Materials. The materials used were identical to those in Experiment 1.

Design and procedure. A 2 (instruction: organize, transcribe) \times 2 (test delay: immediate, delay) between-subjects design was used. The procedures were very similar to Experiment 1 except that in Experiment 2, all of the participants took their notes using a computer and were randomly assigned to the delay groups. After doing the complex span task, the lexical decision task, and lecture note-taking, half of the participants immediately were administered the free recall test followed by the short answer test, whereas the other half were tested 24 hr later. Thus, the participants tested immediately provided a replication of the conditions of current interest from Experiment 1 (i.e., using a computer either to try and transcribe the lecture or to take organized notes).

As in Experiment 1, two independent raters, blind with respect to the conditions, scored all of the notes as well as the free recall responses. Inter-rater reliability was .84 and .91 for notes and free recall responses, respectively, and discrepancies in scoring were resolved by taking the average of the scores given by the two raters.

Results

The groups assigned to the four conditions did not differ in either working memory or processing speed, both $F_s < 1.70$,

precluding the possibility that any group differences in note-taking and test performance could be due to differences in these cognitive abilities.

Note-taking. Note-quantity was greater for those who transcribed compared to those who took organized notes, $F(1, 72) = 24.60$, $p < .001$, $\eta^2 = .26$ (see Table 2). There was no effect of delay and no interaction between the two factors, $F_s < 1.60$.

Free recall. Table 2 presents the mean proportion of total idea units recalled by each group, as well as a breakdown by types of information. There was no effect of instruction on overall free recall, $F < 1.00$, but there was an effect of delay, $F(1, 72) = 23.29$, $p < .001$, $\eta^2 = .22$, indicating that recall was higher when tested immediately as opposed to after a delay. As predicted, there was a delay by instruction interaction, $F(1, 72) = 11.58$, $p < .001$, $\eta^2 = .11$, such that for those instructed to try and transcribe the lecture, performance on the delayed test was significantly poorer than that on the immediate test, $t(36) = 5.20$, $p < .05$, whereas for those instructed to take organized notes, performance did not differ between the immediate and delayed tests, $t < 1.50$. Performance after a delay was better for those who organized compared to those who transcribed, $t(36) = 2.47$, $p < .05$. With respect to main idea units, there was no effect of either instruction or delay, $F_s < 3.13$, but as predicted, there was an interaction, $F(1, 72) = 5.19$, $p < .05$, $\eta^2 = .07$. Although there was no effect of delay for those taking organized notes, $t < 1.00$, those instructed to transcribe recalled less main idea units on the delayed test than on the immediate test, $t(36) = 2.54$, $p < .05$, and their performance was lower than those who organized and were tested at a delay, $t(36) = 3.71$, $p < .01$. With regard to important idea units, there was no effect of instruction or delay, $F_s < 2.80$, but there was an interaction, $F(1, 72) = 8.16$, $p < .01$: There was no effect of delay for those taking organized notes, $t < 1.00$, but those instructed to transcribe showed poorer recall after a delay than when tested immediately, $t(36) = 3.03$, $p < .01$. For unimportant idea units, there was an effect of delay, $F(1, 72) = 12.73$, $p < .01$, but no effect of instruction or an interaction, $F_s < 1.00$.

Short answer. Table 2 shows overall performance on the short answer test by the four groups, as well as a break down into recall of important and unimportant details. There was no main effect of instruction on overall recall, $F < 1.00$, but there was an effect of delay, $F(1, 72) = 13.63$, $p < .01$, indicating that performance on the immediate test was better than performance on the delayed test. There also was an interaction, $F(1, 72) = 10.34$, $p < .001$, such that those who organized did not show a decrement in performance across the delay, $t < 1.00$, whereas those who took transcribed notes did, $t(36) = 5.64$, $p < .01$. As was the case with

Table 2
Experiment 2: Proportions of Idea Units Recalled (Standard Deviations in Parentheses)

Group	Note-taking overall	Free recall				Short answer		
		Overall	Main ideas	Important details	Unimportant details	Overall	Important details	Unimportant details
Immediate								
Organize	.25 (.10)	.12 (.05)	.29 (.12)	.14 (.09)	.12 (.10)	.50 (.19)	.51 (.21)	.44 (.19)
Transcribe	.42 (.15)	.16 (.06)	.30 (.18)	.23 (.12)	.14 (.11)	.64 (.15)	.74 (.19)	.51 (.18)
Delay								
Organize	.25 (.09)	.11 (.04)	.30 (.10)	.17 (.11)	.06 (.07)	.48 (.19)	.51 (.21)	.37 (.19)
Transcribe	.36 (.14)	.07 (.05)	.19 (.09)	.12 (.11)	.05 (.05)	.37 (.15)	.40 (.18)	.28 (.14)

free recall, those who organized showed performance on the delayed test superior to the performance of those who transcribed, $t(36) = 2.02, p < .05$. On questions regarding important details, there was no effect of instruction, $F < 2.00$, but there was an effect of delay, $F(1, 72) = 14.12, p < .001, \eta^2 = .18$, as well as the predicted interaction, $F(1, 72) = 14.20, p < .001, \eta^2 = .18$. When testing was immediate, transcribing led to better performance on an immediate test than on a delayed test, $t(36) = 5.71, p < .05$, but for those instructed to take organized notes, there was no effect of delay, $t < 1.0$. Finally, there was no effect of instruction on recall of unimportant details, $F < 1.00$, but there was an effect of delay, $F(1, 72) = 14.75, p < .01, \eta^2 = .12$, as well as an interaction, $F(1, 72) = 4.09, p < .05, \eta^2 = .03$: There was no effect of delay for those taking organized notes, $t < 1.50$, but those instructed to transcribe showed poorer performance on a delayed test than on an immediate test, $t(36) = 4.58, p < .05$.

Replication. The results for the immediate free recall and short answer tests in Experiment 2 replicated those in Experiment 1: Those who transcribed using a computer had better immediate performance than those who took organized notes on both the free recall and short answer tests, $t(36) = 2.38, p < .05$, and $t(36) = 2.53, p < .05$, respectively.

Discussion

The results of Experiment 2 replicate the finding in Experiment 1 that when notes are taken with a computer, the instruction to transcribe a lecture leads to better immediate test performance than the instruction to take organized notes. However, the pattern of performance reversed after a 24-hr delay. Whereas trying to transcribe led to better performance on immediate tests than taking organized notes, having taken organized notes yielded better performance on delayed tests. This finding is consistent with a levels-of-processing account (Craik & Lockhart, 1972), which predicts better retention of the lecture material for the organize group than the transcribe group because taking organized notes presumably involves deeper and more thorough processing of the lecture information, whereas transcribing requires only a shallow encoding of the information.

The findings of the current experiment may also be conceptualized in terms of Bjork and Bjork's (1992) distinction between storage strength and retrieval strength. This theory assumes that the probability of recalling a target memory depends only on the item's retrieval strength, and that retrieval strength decreases over time. The retrieval strength of an item is mediated by its storage strength, such that items with high storage strength will show less rapid decreases in retrieval strength than items with low storage strength. In Experiment 2, taking organized notes may have resulted in items with higher storage strength than trying to transcribe the lecture, so that these items showed little decline in retrieval strength over the 24-hr delay. Conversely, trying to transcribe the lecture resulted in items with low storage strength that therefore showed a substantial decrease in retrieval strength and considerable forgetting after 24 hr. Further, the finding that, compared to trying to transcribe the lecture, taking organized notes led to poorer immediate learning but superior long-term retention is consistent with the idea of desirable difficulties (Bjork, 1994): Seemingly difficult learning conditions can actually lead to more durable learning. In the present context, it may be assumed that

taking organized notes is more difficult than transcribing what is said, and that this is actually beneficial for long-term retention.

These results suggest that although trying to transcribe a lecture using a computer may be an immediately effective way to take notes, the benefits of such a strategy can be very short-lived. It is relatively uncommon, of course, for students to take lecture notes and then not have an opportunity to study them, as was the case in the first two experiments. It is clear that in order to model more realistic educational scenarios, a third experiment is needed in which the opportunity for students to study their notes is manipulated.

Experiment 3

The preceding experiments establish that transcribing a lecture using a computer can result in better performance on immediate tests than taking organized notes. Presumably, this reflects the benefit of greater note-quantity, although this advantage maybe relatively short-lived. But what about the external storage benefit from note-taking that comes when one studies one's notes? It is unclear which note-taking strategy should lead to a greater external storage benefit because although taking organized notes presumably results in better note quality, trying to transcribe leads to greater note quantity. Experiment 3 pitted note quantity against quality to determine which strategy leads to better learning when, as is typical outside the laboratory, students have the opportunity to study their notes. In Experiment 3, half of the participants were given an opportunity to study their notes and the other half were not. At issue was whether providing a study opportunity would alter the outcome observed in the previous experiment, in which taking organized notes resulted in better performance on delayed tests.

Method

Participants. Seventy-two undergraduate students (47 females and 25 males; mean age = 19.0 years, $SD = 0.9$) at Washington University, all proficient English speakers, participated for course credit.

Materials. The materials used were identical to those used in Experiments 1 and 2.

Design and procedure. A 2 (instruction: organize, transcribe) \times 2 (study: study, no study) design was used. After the lecture, all participants completed the reading span and lexical decision tasks, after which half of the participants were told to study their notes for 5 min. Participants returned 24 hr later for testing. Notes and free recall were scored by two raters blind with respect to the conditions, and discrepancies were resolved by taking the average of the scores. Inter-rater reliability for notes and free recall were .90 and .82, respectively.

Results

As was the case in Experiments 1 and 2, there were no group differences in either working memory or processing speed, both $F_s < 1.31$.

Note-taking. Consistent with the previous experiments, note-quantity was again greater for those who transcribed than for those who took organized notes, $F(1, 68) = 27.48, p < .001, \eta^2 = .29$

(see Table 3). There was no effect of study on note-quantity and no interaction between the two factors, $F_s < 1.00$.

Free recall. Table 3 displays the mean proportion of total idea units recalled by each group, as well as the break down by types of information. There was no effect of instruction on overall recall, $F < 1.00$, but the opportunity to study one's notes did have an effect, $F(1, 68) = 16.50, p < .001, \eta^2 = .12$, which interacted with the note-taking instructions, $F(1, 68) = 9.13, p < .001, \eta^2 = .06$. As in Experiment 2, when participants were not given an opportunity to study, taking organized notes resulted in delayed recall of a higher proportion of total idea units, $t(34) = 2.28, p < .05$. However, when participants were allowed to study, the opposite pattern was observed: Those instructed to transcribe showed better performance than those who took organized notes, $t(34) = 2.11, p < .05$. With respect to main idea units, there was no effect of instruction, nor an effect of study, $F_s < 3.28$. There was an interaction between the two factors, $F(1, 68) = 8.48, p < .001, \eta^2 = .11$: When participants had no opportunity to study their notes, taking organized notes resulted in greater recall than transcribing the lecture, $t(34) = 3.36, p < .05$. However, when participants were allowed to study, there was no effect of note-taking strategy, $t < 1.00$. An effect of instruction on recall of important details was found, $F(1, 68) = 6.51, p < .05$, as well as an effect of study, $F(1, 68) = 4.55, p < .05$, but there was no interaction, $F < 3.13$. Finally, studying led to greater recall of unimportant idea units than not studying, $F(1, 68) = 10.87, p < .01$, but there was no effect of instruction, nor an interaction, $F_s < 2.85$.

Short answer. Table 3 presents performance on the short answer questions. There was no effect of study or instruction on overall recall, $F_s < 3.29$, although there was an interaction, $F(1, 68) = 16.07, p < .001, \eta^2 = .19$: When participants were not allowed to study their notes, performance was better for those who organized compared to those who transcribed, $t(34) = 2.27, p < .05$, whereas when a study period was provided, those instructed to transcribe performed better than those who took organized notes, $t(34) = 3.38, p < .01$. There was no effect of instruction on recall of important details, $F < 1.00$, but the opportunity to study had an effect, $F(1, 68) = 30.19, p < .001, \eta^2 = .26$, which interacted with the instructions, $F(1, 68) = 27.25, p < .001, \eta^2 = .23$. As in Experiment 2, when participants were not allowed to study their notes, taking organized notes led to better performance on a delayed test than transcribing, $t(34) = 2.99, p < .05$. When an opportunity to study one's notes was provided, however, the opposite pattern was observed: Those who tried to transcribe the

lecture showed greater recall of important details, $t(34) = 1.81, p < .05$. Finally, there was no effect of either instruction or study opportunity on recall of unimportant details, $F_s < 1.0$, but there was an interaction, $F(1, 68) = 4.96, p < .05, \eta^2 = .04$: When participants were not allowed to study, there was no difference in recall between those who took organized notes and those who transcribed the lecture, $t < 1.00$, but when participants were allowed to study, those who tried to transcribe the lecture performed better, $t(34) = 2.08, p < .05$.

Discussion

As in Experiment 2, taking organized notes yielded better test performance than trying to transcribe the lecture when tests were given after a 24-hr delay and participants had no opportunity to study their notes. When participants were allowed to study, however, those who had tried to transcribe the lecture were the ones who showed superior delayed recall. These results demonstrate that the benefits of the transcription strategy, which were observed on immediate tests in Experiment 1, can be maintained for at least 24 hr if students are given a brief opportunity to study their notes shortly after the end of a lecture.

Individual differences. The results reported so far provide information regarding the advantages and disadvantages of different note-taking methods and strategies at the group level. A further aim of this study was to examine the role of working memory in note-taking in order to determine what kinds of individuals benefit from these different strategies. If typical note-taking (i.e., taking organized notes) relies on working memory to hold and manipulate lecture information, as Olive and Piolat (2002) hypothesized, then one consequence may be that students with poor working memory are unable to take notes effectively, and thus for these students, studying their notes will provide relatively little benefit. We hypothesize further that simply trying to transcribe a lecture should not require working memory to the same degree as taking organized notes, and therefore students with poor working memory may be able to use this strategy as effectively as those whose working memory ability is much better.

To evaluate these hypotheses, we examined individual differences in working memory and their relation to note-quantity and free recall using data from all three experiments in the present study. To maximize statistical power, we pooled data from similar groups across the three experiments, which yielded four groups of participants who all took notes using a computer. One group took organized notes and was tested immediately, and another group

Table 3
Experiment 3: Proportions of Idea Units Recalled (Standard Deviations in Parentheses)

Group	Note-taking overall	Free recall				Short answer		
		Overall	Main ideas	Important details	Unimportant details	Overall	Important details	Unimportant details
No study								
Organize	.26 (.11)	.12 (.04)	.29 (.15)	.19 (.07)	.07 (.04)	.52 (.18)	.58 (.20)	.43 (.19)
Transcribe	.42 (.14)	.09 (.02)	.16 (.06)	.12 (.06)	.07 (.04)	.41 (.11)	.41 (.15)	.37 (.15)
Study								
Organize	.28 (.12)	.13 (.04)	.25 (.10)	.20 (.07)	.09 (.06)	.49 (.15)	.59 (.14)	.37 (.18)
Transcribe	.45 (.17)	.16 (.06)	.28 (.12)	.18 (.07)	.14 (.07)	.67 (.16)	.81 (.16)	.50 (.19)

took transcribed notes and was also tested immediately; each of these two groups consisted of participants in Experiment 1 and Experiment 2. The other two groups used these same two note-taking strategies but were tested after a 24-hr delay; each of these groups consisted of participants in Experiment 2 and Experiment 3. Finally, we looked at data from Experiment 3, where participants were allowed to study their notes and then were tested 24 hr later.

Immediate testing. Descriptive statistics of the measures used in the individual differences analyses of data from those tested immediately after the lecture are provided in Table 4, with the intercorrelations among these measures presented in Table 5. For both groups, processing speed was a significant predictor of working memory, and note-quantity predicted free recall. Of particular interest, however, are the differences between the two groups. Whereas working memory predicted both note-taking quantity and free recall in the organized note-taking group, working memory was not a significant predictor of either note-quantity or free recall for those who were told to transcribe the lecture.

This absence of a correlation between working memory and either note-quantity or recall of the lecture is unusual in the literature, yet was expected here given our hypothesis that transcribing minimizes the need to hold and manipulate lecture information. Although recall performance did not correlate with working memory, it did correlate with note-quantity, suggesting that students with poor working memory could use the transcription strategy precisely because it relies simply on how fast they can take notes. This is especially important because the results for those taking organized notes indicate that with this latter strategy, those with poor working memory are at a disadvantage when they are given tests that assess their recall of the lecture material. Taken together, our findings suggest not only that transcribing using a computer can lead to superior immediate recall, but also that working memory does not have to play a role in this process.

Delayed testing. Tables 6 and 7 provide descriptive statistics and intercorrelations, respectively, when testing was delayed with no opportunity for participants to study their notes. As expected, processing speed again correlated with working memory regardless of note-taking strategy. More importantly, note-quantity again predicted free recall regardless of note-taking strategy, attesting to the powerful role of sheer note-quantity as a predictor of test performance, regardless of whether testing is immediate or de-

Table 5
Correlations Between Processing Speed, Working Memory, Note-Taking, and Immediate Free Recall

Measure	1	2	3	4
Transcribe condition (<i>n</i> = 38)				
1. Processing speed	1.00			
2. Working memory	-.33*	1.00		
3. Note-quantity	.17	-.05	1.00	
4. Free recall	.24	-.15	.35*	1.00
Organize condition (<i>n</i> = 39)				
1. Processing speed	1.00			
2. Working memory	-.39*	1.00		
3. Note-quantity	-.14	.45*	1.00	
4. Free recall	-.25	.33*	.47*	1.00

* *p* < .05.

layed. With respect to the two note-taking strategies, working memory predicted note-quantity only for those who took organized notes, replicating the results for those tested immediately. The one difference between the results for those tested immediately and those tested after a delay was that in the latter case, working memory and free recall were significantly correlated for both note-taking strategies: Those with higher working memory ability showed less forgetting over the delay. With respect to finding an effective note-taking strategy for those with lower working memory ability, these results may appear problematic. However, they come from participants who were not allowed to study their notes, and as we will show, the situation appears to be different when studying is allowed.

Delayed testing after studying. Experiment 3 demonstrated that although taking organized notes led to better delayed test performance than the transcription strategy when there was no study opportunity, the opposite was true when participants had the opportunity to study their notes—in this case, those using the transcription strategy did better on the delayed tests. For those taking organized notes, both working memory and note-quantity were moderately correlated with free recall after a 24-hr delay (*r* = .30 and *r* = .28, respectively). For those using the transcription strategy, however, note-quantity was strongly correlated with free recall, *r* = .63, but working memory was not, *r* = -.01. Although the sample is obviously too small to draw firm conclusions (*n* =

Table 4
Descriptive Statistics for Processing Speed, Working Memory, Note-Quantity, and Immediate Free Recall

Measure	<i>M</i>	<i>SD</i>	Range
Transcribe condition (<i>n</i> = 38)			
Processing speed	602.3	91.3	448.0
Working memory	35.5	9.0	38.0
Note-quantity	52.7	17.0	69.3
Free recall	20.5	6.9	26.7
Organize condition (<i>n</i> = 39)			
Processing speed	599.6	59.5	242.0
Working memory	35.6	7.4	30.0
Note-quantity	36.5	15.1	69.8
Free recall	14.9	5.9	26.5

Table 6
Descriptive Statistics for Processing Speed, Working Memory, Note-Quantity, and Delayed Free Recall

Measure	<i>M</i>	<i>SD</i>	Range
Transcribe condition (<i>n</i> = 37)			
Processing speed	594.6	73.8	311.0
Working memory	34.5	7.5	32.0
Note-quantity	52.0	18.1	67.0
Free recall	10.7	4.6	23.0
Organize condition (<i>n</i> = 37)			
Processing speed	605.6	87.6	387.6
Working memory	36.8	9.5	39.0
Note-quantity	32.7	32.7	48.0
Free recall	14.2	5.3	22.8

Table 7
Correlations Between Processing Speed, Working Memory,
Note-Taking, and Delayed Free Recall

Measure	1	2	3	4
Transcribe condition ($n = 37$)				
1. Processing speed	1.00			
2. Working memory	-.33*	1.00		
3. Note-quantity	-.16	.05	1.00	
4. Free recall	-.16	.35*	.37*	1.00
Organize condition ($n = 37$)				
1. Processing speed	1.00			
2. Working memory	-.37*	1.00		
3. Note-quantity	.03	.41*	1.00	
4. Free recall	-.12	.36*	.40*	1.00

* $p < .05$.

18), given these results it seems unlikely that increasing the number of observations would lead to a significant correlation between working memory and free recall for those using a transcription strategy.

General Discussion

Three experiments compared the effectiveness of taking organized notes with a transcription strategy in which students try to record as much of a lecture as possible. The results of Experiment 1 revealed that when students took notes by hand and were tested immediately after the lecture, both strategies were equally effective for recall. When students took notes using a computer, however, trying to transcribe the lecture resulted in better test performance than taking organized notes—better, in fact, than using either strategy but taking notes by hand. These results are consistent with the generation effect (Slamecka & Graf, 1978) as well as the translation hypothesis (Conway & Gathercole, 1990), both of which predict that engaging in generative activity during note-taking improves memory.

The results of Experiment 2 revealed that if participants did not study their notes after taking them, the initial advantage that came from use of the transcription strategy with a computer was gone 24 hr later, and those who took organized notes did significantly better. From the perspective of a levels-of-processing framework (Craik & Lockhart, 1972) and Bjork and Bjork's (1992) distinction between storage and retrieval strength, this is not surprising if transcribing information involves shallow processing, whereas organizing information involves deeper, semantic processing, which promotes long-term retention. For participants in Experiment 3 who briefly studied their notes shortly after the lecture and who were tested 24 hr later, transcription was once again the most effective strategy. Because using this strategy with a computer resulted in greater note-quantity than an organized note-taking strategy, we attribute the superior delayed recall of those who used the transcription strategy and then reviewed their notes primarily to an external storage benefit. Taken together, our results suggest that, with respect to the issue of note quality versus quantity, which one plays a more important role for recall depends on the situation. When testing was delayed, for example, those who used the transcription strategy and took more notes did better if studying one's notes was allowed; when studying was not allowed, how-

ever, then those who took higher quality, organized notes did better.

One concern is that participants might not have followed our instructions with respect to note-taking strategy. However, this concern is reduced by examination of the types of idea units that participants recorded in their notes. We would expect to find more main ideas as a proportion of the total number of idea units in the notes of those told to organize compared to those told to transcribe. Indeed, the proportion of main idea units was greater for those told to take organized notes, $t(186) = 2.65$, $p < .001$. Conversely, if those told to try to transcribe the lecture simply typed everything they heard, we should see a greater proportion of unimportant details in their notes compared to those told to take organized notes. This expectation was also confirmed, $t(186) = 4.43$, $p < .001$. Thus, examination of both the quantity and differential selectivity of participants' notes provides no reason to doubt that participants were using the note-taking strategies they were told to.

Researchers have argued that working memory is critical for effective note-taking (e.g., Piolat et al., 2005) because working memory allows individuals to take what is said and (re)organize it into a concise outline of the lecturer's most important points. This process of creating organized notes is reminiscent of certain reading strategies whose goal is to increase comprehension for the material. In such cases, successful text comprehension has been linked to reading strategies such as self-explanation (McNamara, 2004) and generating inferences (McNamara, 2001), which have been thought to be critical for deep-level comprehension of texts (Best, Rowe, Ozuru, & McNamara, 2005). Given the similar goals of strategy use in reading and taking organized lecture notes, it should come as no surprise that there is evidence to suggest that working memory is also related to effective reading strategies (Daneman & Carpenter, 1980).

The relation between working memory and note-taking can create a dire situation for individuals with lower working memory ability who may be unable to take effective organized notes. Previous studies have been inconsistent with respect to the relation between working memory and note-taking (Kiewra & Benton, 1988; Peverly et al., 2007), and as suggested earlier, this may be because different note-taking strategies vary in their reliance on working memory. Indeed, the present results show that whereas taking organized notes depends on working memory ability, the effectiveness with which one can use a transcription strategy does not. For the participants in Experiment 3 who used a transcription strategy and then studied their notes, test performance depended only on note-quantity and not on working memory, suggesting that this strategy may help level the playing field in terms of learning outcomes.

It would be myopic, of course, to introduce the strategy of transcribing notes using computers without discussing potential boundary conditions. To begin with, the benefits of this strategy are undoubtedly limited to those who can type reasonably well. We would note, however, that in the present study, 96% of the participants said they were proficient typists, and we suspect that this is true for college students in general. Another potential boundary condition concerns the extent to which the transcription note-taking strategy improves performance on more conceptually-oriented tests (e.g., those that require reasoning, induction, or transfer). Indeed, it is possible that note-quantity, to the extent that

it reflects shallow processing of the lecture material, could be inversely related to performance on conceptually-oriented tests that may require deeper levels of processing (Bretzing & Kulhavy, 1979; Kiewra, 1985; Kobayashi, 2005; Mayer, 1992). Given that educators are interested in students' grasp of information on a conceptual level, future studies are needed that address this important issue. Another potential concern is that it is unclear whether asking students to try and transcribe an entire lecture is reasonable. Because our lecture passage was only 11 min long, factors such as attention and fatigue may not have played a role, although they might have a much larger influence with longer lectures.

Future studies should explore the extent to which our results apply to e-learning environments, as these may provide a way to deal with this issue of lecture duration. Access to lectures on demand, for example, may allow students to experience a lecture in sections, rather than all at once, if they so choose. This would allow them to select how long they wanted to take notes for before reviewing the material, thereby potentially maximizing the benefits of the transcription note-taking strategy. Finally, it is not immediately clear how the current findings would apply to lectures in different disciplines, particularly those in which material is presented graphically, diagrammatically, or pictorially. Nevertheless, even as the variety of educational experiences available is expanding such that students are able to attend lectures from home via the Internet, or to develop a schedule of learning activities that best suits them, the role of note-taking is likely to remain critical for creating meaningful information from lecture material.

Instructors spend most of their class time lecturing, and students need effective note-taking skills and strategies in order to do well on exams. However, exactly what constitutes the most effective note-taking strategies may vary across students who differ in cognitive ability. As a result, research is needed on how individual differences interact with note-taking strategies so that students can be guided toward strategies that rely on cognitive abilities that they are stronger in, or toward strategies that depend less on the abilities that they are weaker in. For students who have poor working memory, at least, help may be on its way.

References

- Armbruster, B. B. (2000). Taking notes from lectures. In R. F. Flippo & D. C. Caverly (Eds.), *Handbook of college reading and study strategy research* (pp. 175–199). Hillsdale, NJ: Erlbaum.
- Baddeley, A. D. (1986). *Working memory*. Oxford, England: Oxford University Press.
- Best, R. M., Rowe, M., Ozuru, Y., & McNamara, D. S. (2005). Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders*, 25, 65–83. doi:10.1097/00011363-200501000-00007
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.
- Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp.35–67). Hillsdale, NJ: Erlbaum.
- Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. *Contemporary Educational Psychology*, 4, 145–153. doi:10.1016/0361-476X(79)90069-9
- Brown, C. M. (1988). Comparison of typing and handwriting in “two-finger typists.” *Proceedings of the Human Factors Society, USA*, 32, 381–385.
- Carnes, M. C. (1995). *Past imperfect: History according to the movies*. New York, NY: Holt.
- Cohn, E., Cohn, S., & Bradley, J. (1995). Notetaking, working memory, and learning in principles of economics. *The Journal of Economic Education*, 26, 291–307. doi:10.2307/1182993
- Conway, M. A., & Gathercole, S. E. (1990). Writing and long-term memory: Evidence for a “translation” hypothesis. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 42, 513–527.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684. doi:10.1016/S0022-5371(72)80001-X
- Crawford, C. C. (1925a). The correlation between lecture notes and quiz papers. *The Journal of Educational Research*, 12, 282–291.
- Crawford, C. C. (1925b). Some experimental studies on the results of college note-taking. *The Journal of Educational Research*, 12, 379–386.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466. doi:10.1016/S0022-5371(80)90312-6
- Di Vesta, F. J., & Gray, G. S. (1972). Listening and note taking. *Journal of Educational Psychology*, 63, 8–14. doi:10.1037/h0032243
- Dunkel, P., & Davy, S. (1989). The heuristic of lecture notetaking: Perceptions of American & international students regarding the value & practice of notetaking. *English for Specific Purposes*, 8, 33–50. doi:10.1016/0889-4906(89)90005-7
- Faber, J. E., Morris, J. D., & Lieberman, M. G. (2000). The effect of note taking on ninth grade students' comprehension. *Reading Psychology*, 21, 257–270. doi:10.1080/02702710050144377
- Fisher, J. L., & Harris, M. B. (1973). Effect of note taking and review on recall. *Journal of Educational Psychology*, 65, 321–325. doi:10.1037/h0035640
- Kiewra, K. A. (1985). Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist*, 20, 23–32. doi:10.1207/s15326985ep2001_4
- Kiewra, K. A., & Benton, S. L. (1988). The relationship between information processing ability and notetaking. *Contemporary Educational Psychology*, 13, 33–44. doi:10.1016/0361-476X(88)90004-5
- Kiewra, K. A., Benton, S. L., & Lewis, L. B. (1987). Qualitative aspects of notetaking and their relationship with information-processing ability and academic achievement. *Journal of Instructional Psychology*, 14, 110–117.
- Kiewra, K. A., DuBois, N. F., Christensen, M., Kim, S., & Lindberg, N. (1989). A more equitable account of the note-taking functions in learning from lecture and from text. *Instructional Science*, 18, 217–232. doi:10.1007/BF00053360
- Kobayashi, K. (2005). What limits the encoding benefit of note-taking? A meta-analytic examination. *Contemporary Educational Psychology*, 30, 242–262. doi:10.1016/j.cedpsych.2004.10.001
- Mayer, R. E. (1992). Cognition and instruction: Their historic meeting within educational psychology. *Journal of Educational Psychology*, 84, 405–412. doi:10.1037/0022-0663.84.4.405
- McNamara, D. S. (2001). Reading both high-coherence and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, 55, 51–62.
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38, 1–30. doi:10.1207/s15326950dp3801_1
- Olive, T., & Piolat, A. (2002). Suppressing visual feedback in written composition: Effects on processing demands and coordination of the

- writing processes. *International Journal of Psychology*, 37, 209–218. doi:10.1080/00207590244000089
- Palmatier, R. A., & Bennett, J. M. (1974). Notetaking habits of college students. *Journal of Reading*, 18, 215–218.
- Peverly, S. T., Ramaswamy, V., Brown, C., Sumowsky, J., Alidoost, M., & Garner, J. (2007). What predicts skill in lecture note taking? *Journal of Educational Psychology*, 99, 167–180. doi:10.1037/0022-0663.99.1.167
- Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology*, 19, 291–312.
- Rabinowitz, J. C., & Craik, F. I. M. (1986). Specific enhancement effects associated with word generation. *Journal of Memory and Language*, 25, 226–237. doi:10.1016/0749-596X(86)90031-8
- Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology*, 97, 70–80. doi:10.1037/0022-0663.97.1.70
- Rogers, J., & Case-Smith, J. (2002). Relationship between handwriting and keyboarding performance of sixth-grade students. *American Journal of Occupational Therapy*, 56, 34–39. doi:10.5014/ajot.56.1.34
- Scardamalia, M., & Bereiter, C. (1986). Research on written composition. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 778–803). New York, NY: Macmillan.
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 592–604. doi:10.1037/0278-7393.4.6.592
- Smith, S. D., & Caruso, J. B. (2010). *The ECAR Study of Undergraduate Students and Information Technology, 2010* (Research Study, Vol. 6). Available from <http://www.educause.edu/ecar>
- Williams, R. L., & Eggert, A. C. (2002). Notetaking in college classes: Student patterns and instructional strategies. *The Journal of General Education*, 51, 173–199. doi:10.1353/jge.2003.0006
- Wirt, J., Choy, S., Greal, D., Provasnik, S., Rooney, P., & Watanabe, S. (2001). *The condition of education* (NCES Publication No. 2001072). Washington, DC: Government Printing Office.

Received September 22, 2011

Revision received August 23, 2012

Accepted September 4, 2012 ■

Call for Papers: Special Issue Ethical, Regulatory, and Practical Issues in Telepractice

Professional Psychology: Research and Practice will publish a special issue on recent ethical, regulatory and practical issues related to telepractice. In its broadest definition the term telepractice refers to any contact with a client/patient other than face-to-face in person contact. Thus, telepractice may refer to contact on a single event or instance such as via the telephone or by means of electronic mail, social media (e.g., Facebook) or through the use of various forms of distance visual technology. We would especially welcome manuscripts ranging from the empirical examination of the broad topic related to telepractice to those manuscripts that focus on a particular subset of issues associated with telepractice. Although manuscripts that place an emphasis on empirical research are especially encouraged, we also would welcome articles on these topics that place an emphasis on theoretical approaches as well as an examination of the extant literature in the field. Finally, descriptions of innovative approaches are also welcome. Regardless of the type of article, all articles for the special issue will be expected to have practice implications to the clinical setting. Manuscripts may be sent electronically to the journal at <http://www.apa.org/pubs/journals/pro/index.aspx> to the attention of Associate Editor, Janet R. Matthews, Ph.D.

A (Pan-Canadian) Cluster Randomized Control Effectiveness Trial of the ABRACADABRA Web-Based Literacy Program

Robert Savage
McGill University

Philip C. Abrami
Concordia University

Noella Piquette
Lethbridge University

Eileen Wood
Wilfrid Laurier University

Gia Deleveaux
Concordia University

Sukhbinder Sanghera-Sidhu and Giovani Burgos
McGill University

This report describes a cluster randomized control trial (RCT) intervention study of the effectiveness of the ABRACADABRA (ABRA) Web-based literacy system using a classroom-level RCT intervention with 1,067 children in 74 kindergarten and Grade 1 or Grade 1/2 classrooms across Canada. The authors closely followed the CONSORT criteria for executing and reporting high-quality RCT studies. Well-trained teachers delivered the ABRA intervention to their regular classrooms for 20 hr per child over one full semester. At posttest, the ABRA intervention classroom showed significant advantages over controls in phonological blending ability, letter-sound knowledge and, marginally, for phoneme segmentation fluency. A secondary analysis exploring the effects of different levels of program implementation showed that with fidelity of implementation (80% of intervention teachers), advantages were evident at posttests in phonological blending, phoneme segmentation fluency, sight word reading, and letter-sound knowledge. It is concluded that ABRA is an effective resource for key skills associated with early reading. Implications for the role of both Web-based technologies and extended professional development for technology in aiding in the scale-up of evidence-based reading interventions are discussed.

Keywords: reading, intervention, technology, randomized control trial

What evidence is there that technology can aid reading acquisition? Up until very recently, it could be fairly remarked that much enthusiasm exists for the potential of computer-based information and communication technologies (ICT) for helping children in learning to read (Bereiter, 2002; Dede, 1996; Harasim, Hiltz, Teles, & Turoff, 1995; Mayer, 2001; Rabiner & Malone, 2004; Scardamalia & Bereiter, 1996). There is also widespread deployment of ICT in North American schools (Chambers,

Abrami, et al., 2008; Cuban, 2001). Nevertheless, expectation in this domain has run far in advance of the existing evidence base (Andrews et al., 2007; Chambers, Abrami, et al., 2008; Savage, Abrami, Hipps, & Deault, 2009). Beyond the effectiveness of technology, there is also a need to closely tie ICT use to contemporary theoretical models of reading acquisition and to coherent pedagogical models for technology.

Turning first to the evidence base for technology and reading, several reviews of research that have included quasi-experimental studies have identified small effect sizes for ICT on literacy, and researchers are therefore cautiously optimistic about ICT (Blok, Oostdam, Otter, & Overmatt, 2002; Ehri et al., 2001; MacArthur, Ferretti, Okolo, & Cavalier, 2001). However, to confidently answer the question, “Does technology cause improvement in reading?” methodologically, high-quality randomized control trials (RCTs) must occupy a central role, as they provide unique protection that effects reported are due to the intervention. Finally, systematic reviews of the whole literature including *only* such studies provide the greatest confidence in findings (Savage, 2012; Torgerson, Brooks, & Hall, 2006).

In this specific methodological sense then, the most reliable evidence concerning ICT can be derived from three comprehensive systematic reviews (Andrews et al., 2007; Slavin, Cheung, Groff, & Lake, 2008; Slavin, Lake, Chambers, Cheung, & Davis, 2009; Torgerson & Zhu, 2003). Torgerson and Zhu (2003) found

This article was published Online First December 17, 2012.

Robert Savage, Department of Educational and Counseling Psychology, McGill University, Montreal, Quebec, Canada; Philip C. Abrami, Centre for the Study of Learning and Performance, Concordia University, Montreal, Quebec, Canada; Noella Piquette, Faculty of Education, Lethbridge University, Lethbridge, Alberta, Canada; Eileen Wood, Department of Developmental Psychology, Wilfrid Laurier University, Waterloo, Ontario, Canada; Gia Deleveaux, Centre for the Study of Learning and Performance, Concordia University; Sukhbinder Sanghera-Sidhu, Department of Educational and Counseling Psychology, McGill University; Giovani Burgos, Department of Sociology, McGill University.

Correspondence concerning this article should be addressed to Robert Savage, Department of Educational and Counseling Psychology, Faculty of Education, McGill University, Montreal H3A 1Y2, Canada. E-mail: robert.savage@mcgill.ca

only 12 RCT studies among an initial 2,319 studies that examined ICT-based interventions for children ages 5–16 years. Torgerson and Zhu found that across these 12 studies, some effect sizes were positive, and some, including those of the largest RCT study, were actually negative. Overall, only small and nonsignificant effect sizes for interventions were evident. Torgerson and Zhu (2003) concluded that teachers should not use technology to teach literacy until several, new, well-designed RCT studies with consistently positive effects have been published and evaluated (see also Torgerson, 2007).

Similarly, a systematic review of technology on middle and high school learning by Slavin et al. (2008) found few ($n = 8$) well-designed intervention studies. The mean overall effect size was a very modest $+0.10$, suggesting that computers do not aid literacy skill development in any substantial way. Slavin et al. (2009) also reported similar patterns for 12 elementary school studies. It is important to note that Slavin et al. reviewed only published articles and did not seek unpublished reports, which may inflate findings due to a publication bias. Their studies also included quasi-experiments in addition to RCT studies.

Similar trends are evident in the much of the recent research literature. Using a large-scale RCT study design, Dynarski et al. (2007) evaluated technology-based reading interventions using a range of commercially available ICT products in 132 schools with a total of 4,389 teachers. Dynarski et al. found that the mean effect size for interventions was not significantly different from zero on any tests of reading in either Grade 1 or Grade 4. In a follow-up study, Campuzano, Dynarski, Agodini, Rall, and Pendleton (2009) explored a representative subset of classrooms from the Dynarski et al. study and again reported almost no significant effects of technology intervention on attainment and no effects on literacy whatsoever.

Finally, Andrews et al. (2007) carried out a systematic review of the impact of technology on written language learning. They identified only eight reasonably well-designed intervention studies from an initial search of 2,103 recent articles and reports, of which only one of the included studies (Lewis, Ashton, Haapa, Kieley, & Fielden, 1999) had used some random allocation of participants. Andrews et al. concluded that while many studies purport to have demonstrated effectiveness of ICT, none have done so convincingly. Andrews et al. also drew attention to the lack of clear definition of terms and the absence of theorizing of the role of technology in supporting literacy. They summed up the state of the literature on ICT as follows:

[T]he field is in a preparadigmatic state where definitions of English, literacy, and ICT are still relatively unclear and where the causal or reciprocal relationships between them have yet to be fully theorized. (p. 325)

Methodological Issues in Interpreting Effects of ICT

In evaluating the pessimistic findings of this RCT-based literature, there are perhaps three overarching methodological issues that should be noted. The first issue concerns *implementation* of studies, the second issue concerns the *quality of the technology* used in research studies, and a final issue concerns what we term the *theoretical and pedagogical coherence* of technologies and their implementation. Turning first to implementation, teacher-led reading intervention research suggests that the quality of the im-

plementation of programs has a significant impact on outcomes both when the interventions involve technology (Chambers, Abrami, et al., 2008; Savage et al., 2010; Wolgemuth et al., 2011) and when they do not involve technology (e.g., Davidson, Fields, & Yang, 2009; Lane, Bocian, MacMillan, & Gresham, 2004; Stein et al., 2008). Indeed, in several recent studies, variation in program implementation was the biggest factor moderating outcomes and could explain *entirely* the implementation effects reported (e.g., Chambers, Abrami, et al., 2008; Davidson et al., 2009). Rigorous and detailed investigation of the ways teachers implement (or do not implement) interventions clearly needs to be explored in detail in ICT studies, yet this is rarely reported. In the Campuzano et al. (2009) study, for example, there was no assessment of treatment integrity whatsoever. This issue of teachers' fidelity to treatment protocols is particularly problematic in ICT-based interventions. Much research shows that a host of teacher factors—including teachers' use of ICT at home (Wozney et al., 2006); comfort with computers (Chen & Chang, 2006); beliefs and expertise (Chen, 2008; Sang, Valcke, van Braak & Tondeur, 2010), and fear and other emotional reactions toward computers—affect teachers' actual use of computer technology (see Kay, 2008; Mueller, Wood, & Willoughby, 2008; Wood, Specht, Willoughby, & Mueller, 2008).

Some formal models of teacher's use of technology posit stages of development (e.g., Mueller, Wood, Willoughby, Ross, & Specht, 2008; Sandholtz, Ringstaff, & Dwyer, 1997). In the teachers' use of technology model by Sandholtz et al., the *entry* level is characterized by the making of time-consuming mistakes, frustration, and high levels of discontinuation. Teachers at the *adoption* level can use technology in a systematic manner. Adoption is characterized adherence to treatment protocols (treatment integrity) with minimum experimentation or formation of links to other forms of learning, such as collaborative or experientially based learning. Adaptation, by contrast, steps beyond conventional treatment integrity. At the *adaptation* level, teachers' integration of technology has been described as transforming classroom teaching, with greater connectivity between all forms of learning. Savage et al. (2010) found that evidence of all three of these types of practice in the use of an Internet-based reading software program (ABRACADABRA, A Balanced Reading Approach for Canadians Designed to Achieve Best Results for All). Furthermore, teachers rated as being at the adaptation level produced significantly greater student gains in reading than teachers at other levels of practice. Wolgemuth, Helmer, Bottrell, Harper, and Lea (2012) reported comparable effects of implementation quality on student outcomes. In another model, Wozney, Venkatesh, and Abrami (2006) used expectancy theory to examine teachers' use of technology related to three broad motivational categories: perceived *expectancy* of success, perceived *value* of technology use, and perceived *cost* of technology use. Each category impacted on use. Expectancy beliefs mattered the most; value factors were positively associated with use, but cost factors were negatively associated with use.

A closely related issue concerns the training given for technology use in intervention research. Teachers are unlikely to successfully implement any program if they have limited training and support in its use in the classroom. It is nevertheless quite typical in ICT studies to offer a single day of training despite the widespread understanding that such models of professional learning

produce extremely limited professional change (e.g., Clarke & Hollingsworth, 2002; Fullan, 2001; Waks, 2007) and are widely discredited “straw man” models. The Dynarski et al. (2007) study was quite typical in this regard: Teachers were given a single-day training session on the technology several weeks before the intervention commenced. Teacher self-reports showed that around a third of teachers did not feel qualified to use the technology at the start of the intervention. As noted earlier, the observed quality of classroom delivery of ICT was not explored, so the true capacity of teachers to deliver the trained interventions is unknown. The quality of teacher training has also not been reported as a factor in systematic reviews of ICT effectiveness.

It is possible, however, to effectively support teachers in their use of new technology. Anderson, Wood, Piquette-Tomei, Savage, and Mueller (2011) documented all requests from 10 teachers in kindergarten and Grades 1 and 2 who received “just-in-time” instructional support over a 2-½-month period while implementing a novel technology-based intervention. From detailed observations, subsequent analysis of 80 just-in-time support sessions indicated that the greatest level of support was required during the initial stages of implementation and for technological rather than pedagogical questions. Support requests declined sharply over the first 3 weeks or so of implementation, and most of the problems were resolved immediately through the just-in-time support system. These results inform the kinds of ongoing supports teachers need to run technology-based interventions effectively.

The second issue is that whether ICT “works” will always be subject to the quality of ICT available. In many studies, experimenters rightly attempt to pick the best technologies available. Dynarski et al. (2007) selected ICT that, according to the manufacturers, had first been research validated. There was, however, no prior independent assessment of the nature and quality of the technology programs used. Recently, Grant et al. (2012) developed a taxonomy of cognitive reading skills based upon both systematic reviews of evidence (e.g., National Reading Panel, 2000) and an expert panel. Grant et al. then used the taxonomy to explore how well 30 commercially available software programs designed to teach early literacy support these key reading skills. Results showed that although some skills such as alphabetic knowledge were being trained in a somewhat developmentally appropriate manner, very few programs (15%) taught synthetic phonics, and none at all taught phoneme segmentation or concepts about print. Key comprehension skills, such as summarizing a text or generating questions, were entirely absent. Overall Grant et al. noted that there were limited examples for training each skill, inconsistent progression, and few opportunities to practice skills. The study by Grant et al. suggests that better evidence-based ICT is needed if ICTs are to be used to assist in reading skill development.

A related final issue that we would draw attention to in evaluating technology is the link between contemporary theories of reading developed inside and outside technology and the pedagogical theory of implementation of ICT. Among the best of recent work in this field are the reports by Chambers, Abrami, et al. (2008) and Chambers, Slavin, et al. (2008) on a series of well-designed RCT studies. Each study found significant positive effects of a technology on literacy used to complement the nontechnology-based Success for All program. Alongside this empirical evidence, Chambers, Slavin, et al. (2008) also explicitly identified a pedagogical role for technology: They articulated what

computers cannot do and delineated a role for the technology as adding value to regular classroom teaching in what they refer to as *embedded ICT*. In addition, Chambers et al. drew upon contemporary theory about how ICT supports reading through encouraging dual visual and verbal coding (Clark & Paivio, 1991) and through “offloading” between modalities to reduce working memory load, (Solso, 2001) thereby encouraging retention.

Beyond these examples, it is rare in ICT studies on literacy to consider contemporary cognitive models of reading. To take certain aspects of word recognition as an example, all current accounts of word reading assume strong connectivity between representations of words or sublexical letter strings (Bishop & Snowling, 2004; Coltheart, Curtis, Atkins, & Haller, 1993; Seidenberg & McClelland, 1989). A key issue thus concerns the specificity of spelling representations in these distributed lexicons. In dominant-phase-based models, partly specified phonologically based cues in words are developmentally important (e.g., Ehri, 2005; Perfetti, 2007; Savage & Stuart, 2006; Share, 1995). Phonological awareness is assumed to be the cutting edge of word learning (Hulme, Snowling, Caravolas, & Carroll, 2005; Share, 1995) and must be applied fluently to word recognition in text. This analysis suggests the need for progression within ICT activities that reflects these qualitative changes in development (see also Ecalte, Magnan, & Calmus, 2009, for theoretically driven research on ICT).

Naturally, there are also a number of unresolved issues. One example is the relative role of rhyme-based inference and use of grapheme-phoneme rules in early reading (see Savage, Deault, Daki, & Aouad, 2011, for a recent review). Another is the relative efficacy of different kinds of phonics programs such as “synthetic phonics” and “analytic phonics” (Torgerson et al., 2006). Wise implementation of ICT in this uncertainty would allow *all* candidate models to be further tested fairly; yet of the 30 ICT products reported by Grant et al. (2012), nearly all implemented only one model of phonics. For all of these reasons, we have over several years developed the ABRACADABRA reading intervention as a means to address previous shortcomings in literacy ICT programs.

The ABRACADABRA Web-Based Literacy Programmatic Research

As mentioned earlier, ABRACADABRA (hereafter, ABRA), is an acronym for A Balanced Reading Approach for Canadians Designed to Achieve Best Results for All. ABRA is a free-access Web-based literacy software for beginning readers. ABRA has four modules: letters and sounds, reading, comprehension, and writing/spelling. It can be used flexibly. ABRA does not need to be used within the context of a particular nontechnology reading program. It contains 32 distinct leveled activities and 17 stories. ABRA was developed by a multiuniversity team (see Abrami et al., 2010; Abrami, Savage, Wade, Hipps, & Lopez, 2008; Hipps, Abrami, Savage, Cerna, & Jorgensen, 2005, for details). ABRA is based on the evidence from systematic reviews on effective reading interventions as well as developmental theory (see Savage et al., 2009, for a review) and includes instructional, professional, and parent modules (see <http://abralite.concordia.ca>). A version of ABRA containing an assessment and communication module is also freely available; however, due to the fact that it can store

student records, this must be downloaded and stored on a school board server (see <http://grover.concordia.ca/abracadabra/>).

ABRA is supported by evidence from three RCT efficacy studies. In a within-classroom RCT design, Comaskey, Savage, and Abrami (2009) sampled 53 children from a disadvantaged low socioeconomic (SES) urban kindergarten where the majority of students experienced English as a second language. Children were exposed to ABRA instruction in small groups for 10 hr over 13 weeks. Phoneme-based teaching led to significant growth in phoneme blending in the ABRA group only. Rhyme-based teaching led to similarly specific growth in rhyme awareness. Di Stasio, Savage, and Abrami (2012) followed up these children 1 year later. They reported that the analytic phonics group performed better on a passage-reading comprehension task than the synthetic phonics group. An overall effect size of 0.41 per hour of intervention was reported.

Savage et al. (2009) also explored the effectiveness of ABRA in 144 typically developing Grade 1 children using a within-classroom RCT design and also contrasted phoneme-based “synthetic” phonics and a rhyme-based “analytic phonics” methods as well as a regular classroom control. Children received the intervention in small groups for an average of 13 hr per child delivered by trained research assistants. Significant differences between treatment and control groups at posttest were evident on phonological awareness, listening comprehension and reading fluency (favoring synthetic phonics), and letter-sound knowledge (favoring analytic phonics). In addition, overall effect sizes were significantly different from zero at immediate and delayed (8 months) posttest after the completed intervention phase in both interventions. In addition, Deault, Savage, and Abrami (2009) explored response characteristics of students in this sample and found that ABRA can moderate the associations between literacy and attention and may support students at risk of reading and attention difficulties.

Finally, Wolgemuth et al. (2012) reported a within-class RCT evaluating ABRA in the Northern Territories in Australia with over 300 children, including a large number of aboriginal students for whom English was an additional language. Specially recruited and trained teachers delivered ABRA as a pull-out program in schools for around 40 min four times a week for 16 weeks. At posttest, the children who received the ABRA intervention showed significant advantages in phonological awareness ($d = 0.37$) and grapheme-to-phoneme knowledge ($d = 0.37$). Aboriginal students made at least the same amount of progress as nonaboriginal students and, sometimes, greater progress.

The Current Study

All current scientific evidence regarding ABRA comes from *efficacy trials*—trials run by university-based researchers and assistants or delivered by specially trained professionals. However, ABRA was developed as a free tool for teachers and placed on the worldwide Web to allow maximal accessibility and scalability of interventions (e.g., Abrami et al., 2008). The next step necessary to establish the impact of this ICT software is large-scale *effectiveness trials* to explore the impact of interventions under real-world conditions undertaken by well-supported regular classroom teachers within their regular classrooms. In such approaches, classrooms, not students, are treated as the unit of analysis. Well-trained

teachers with the ability to implement ICT interventions thus have a central role in such designs. As there is much evidence of the impact of variation in teachers’ delivery of programs, a secondary question concerns the impact of adequate adoption of technology over inadequate adoption and of the possible role of adapted implementation of ABRA.

Thus, there are two main research questions of this article: (a) Does a classroom-level-cluster RCT effectiveness trial of ABRA yield significant advantages for intervention over control classrooms in early literacy at posttest? and (b) Do adequately implementing classrooms and transformational implementations in the classrooms yield additional effects for ABRA?

Method

Design

This study is a cluster RCT intervention study that took place over two academic years, 2007–2008 and 2008–2009. This study used a pretest–posttest experimental intervention design and randomized 74 participating classrooms containing 1,067 students) across the three provinces of Alberta, Ontario, and Quebec, Canada. Randomization took place within schools at the classroom level. Pairs of classrooms were identified within each school and were then randomly allocated to either ABRA intervention ($n = 37$ classrooms) or control non-ABRA regular classroom teaching conditions ($n = 37$ classrooms). Such prior identification reduces bias in cluster RCT designs (Puffer, Torgerson, & Watson, 2005). Randomization was achieved by blindly pulling slips of paper containing teachers’ names from a hat in each school.

We made a formal power calculation using Power and Precision Version 3, (Borenstein, Rothstein, Cohen, Schoenfeld, & Berlin, 2001). This showed that with α set at .05, $n = 35$ cases in each cell, and two covariates, this design would yield a .88 power to detect a large effect size in standard analysis of covariance (ANCOVA, where a large effect is defined, according to Cohen, 1988, as $f = 0.40$). This level of power appeared very reasonable given that Savage et al. (2009) reported large effect sizes for ABRA on standardized measures of phonological awareness.

Students in the control classrooms continued to receive their regular English language arts (ELA) lessons that did not include ABRA, while those in the experimental group had ABRA integrated into their ELA lessons. ABRA was part of regular classroom teaching and was not a supplementary program. Inclusion of ABRA did not result in extra pedagogical time being allocated to literacy in the intervention classrooms. Beyond the ABRA intervention, all teachers in both study conditions were encouraged to carry on ‘business as usual’ in their teaching. The conceptual and methodological issues behind this request were explained to teachers and often revisited. Intervention teachers implemented the ABRA Web-based literacy program 2 hr per week. Posttesting took place after 10–12 weeks of intervention.

Sample and Participant Selection

Classroom teacher recruitment. Starting in the spring of 2007, members of the research team made demonstrations of the ABRA software to principals, teachers, ELA consultants at school

boards, and other educators at various meetings throughout Quebec, Ontario, and Alberta. School administrators and school board officials who then expressed interest received letters asking for their participation. These letters outlined expectations for involvement. Principals of participating schools were required to allow random allocation of classrooms to the intervention or control condition, to ensure that the experimental classes would be guaranteed 2 hr of computer time each week for ABRA implementation, and to allow teachers assigned to the experimental condition to attend ABRA training sessions. In return, participating schools received free ABRA teacher technological and pedagogical training and support. Control classroom teachers received ABRA training after the intervention was completed.

Schools approached for this study were nonselective urban, suburban, and rural elementary schools. From the recruitment presentations, 81 teachers from 24 schools across the three participating provinces responded positively. However, from the group of 81 classrooms that participated in the study, seven of the classrooms involved did not meet the random assignment requirements to be part of the final RCT analysis. Field notes on the recruitment and selection process of the participating classrooms indicated that three of the classroom teachers were not willing to be randomly assigned. Data collected on a delayed intervention group from Ontario also were removed, as this classroom had not been matched to a control comparison group. Two additional intervention groups were removed: a Grade 1–2 split class from Ontario and a Grade 1 class from Quebec, as neither class had been matched with a control comparison classroom within their respective schools. Similarly, data from a kindergarten control class from Quebec were removed, as this class had not been matched with an intervention comparison group. The final ABRA RCT data set contained 74 matched classrooms, consisting of 12 paired kindergarten classrooms ($n = 24$) and 25 paired first- and second-grade classrooms ($n = 50$). The details of the final sample analyzed are specified in the Enrollment and Allocation sections of the CONSORT flow diagram (see Figure 1).

Student participants. Once classroom teachers had agreed to participate in the ABRA study, investigators obtained permission for their students to participate. There were 1,068 students whose parents agreed to allow their child(ren) to participate in the study. One consented kindergarten student moved away before pretesting, which brought the sample total to 1,067 students. The child sample included 316 kindergarten students ($n = 154$ ABRA, $n = 162$ control), 616 first graders ($n = 352$ ABRA, $n = 264$ control), and 135 second grade students ($n = 43$ ABRA, $n = 92$ control). By gender, the final sample of children consisted of 543 girls and 524 boys. The random allocation of classrooms resulted in 284 girls and 265 boys in the experimental group (total $n = 549$), and 259 girls and 259 boys in the control group (total $n = 518$). Age of the participating students at pretest averaged 73.69 months old ($SD = 10.11$ months), with a range of 67 months. No student was excluded due to language or exceptionalities.

The demographic profile of schools was evaluated through a formal report created the Canadian Institute for Social Policy (CRISP; D. Willms, personal communication, April 22, 2011). Here, individual student-level postal code data for all schools in the study ($n = 24$ schools) were linked to data from the long form of the 2006 Canadian Census (Statistics Canada, 2006) to construct a set of derived variables that describe the SES characteristics of

A CONSORT E-flowchart at the classroom level for the Pan-Canadian ABRA Cluster RCT study.

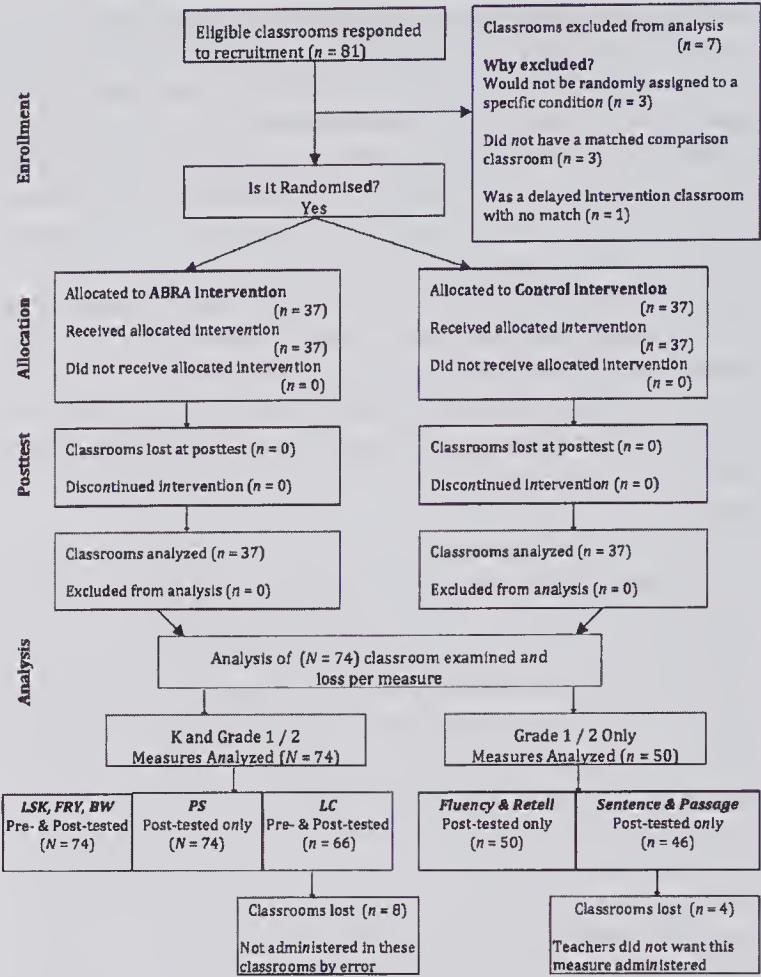


Figure 1. Consolidated Standards of Reporting Trials (CONSORT) E-flowchart showing classroom level for the cluster randomized clinical trial study of the Pan-Canadian ABRA (acronym for ABRACADABRA, A Balanced Reading Approach for Canadians Designed to Achieve Best Results for All). LSK = Letter-Sound Knowledge; FRY = Fry Words; BW = Blending Words subtest; PS = Phonemic Segmentation Fluency subtest; LC = Listening Comprehension measure.

the postal code area in which each child resides. A composite z-standardized measure was derived on the basis of occupational status, unemployment rates, family income, and years of education with national mean of 0 and standard deviation of 1. The study sample scores were normally distributed, with a mean of 0.03 ($SD = 0.55$). This confirmed that schools in our sample closely reflected national school SES patterns.

Procedure

ABRACADABRA sessions. Before intervention, teachers began implementing ABRA in their classrooms; a full day of training took place led by three of the research investigators. Teachers were exposed to the philosophical, developmental, and pedagogical underpinnings of the software and were given hands-on time to explore the software. Theoretically based developmental progression was emphasized throughout training. For example, teachers were shown how the ABRA phonic activities follow theoretically prescribed patterns of expected difficulty (e.g., detection tasks before production tasks, two-phoneme blends (e.g., *a-t*) progressing up sequentially to six-phoneme blending tasks (e.g., *s-p-r-i-*

n-t), and the early emergence of boundary consonants over medial vowels in word recognition and phonological tasks, the asymmetric later introduction of segmenting tasks, introduction of singleton letter sounds before complex digraphs, and so on). As the teachers became more familiar with ABRA during the session, they were encouraged to interact with the various online activities and to plan for student progression. Teachers were made aware that ABRA is only a tool, not an ICT “magic bullet.” They were shown that ABRA requires skilled teachers to implement it well and to link it effectively to cross-curricular learning outside the ABRA sessions. The concept of adaptive transformational use of technology as described earlier and reported in Savage et al. (2010) was discussed explicitly with all intervention teachers.

Once the intervention teachers had some hands-on exposure to the program, the investigators then presented and reviewed a suggested format for the teachers to use during a 1-hr ABRA lesson that specified 10 min of word-level work, 10 min of text-level work, 20 min of collaborative work, and 20 min of extension activities. The word-level work involved activities such as letter knowledge, phonological awareness, phonics, and word-building. Text-level work invites use of the fluency and comprehension activities based on the digital stories component in ABRA. For reading fluency, activities such as high-frequency words, reading with expression, reading accurately, and choral reading were suggested. For comprehension, activities that focused on predicting, comprehension monitoring, recognizing story elements, and summarizing as well as vocabulary and writing were identified. Emphasis was thus placed on demonstrating that ABRA acts as a balanced literacy program (e.g., Pressley, 1998).

All teachers were encouraged to use these activities but to select activities relevant to the appropriate point in children’s development. Thus, for kindergarten teachers, the phonic activities might be more likely to start with simpler activities within ABRA such as sound awareness, syllable and word counting, and aspects of rhyme awareness, whereas the teachers in Grade 1 might move their children more quickly to the simpler of the explicit blending tasks. Similarly for comprehension and aspects of the fluency tasks, teachers of kindergarten children were asked to encourage children to listen to stories and then complete tasks such as story ordering and summarizing, whereas teacher of children in Grade 1 were asked, where appropriate, to encourage children to read the texts and then complete comprehension tasks. This differentiated use of ABRA was also encouraged through the sustained in-class follow-up support for teachers.

Collaborative work encouraged students to work together in order to practice skills they would have learned in the earlier two sections. Collaborative work did not have to be conducted on a computer. This became an opportunity for students to work with and learn from each other. For example, we suggested that students could write alternative endings for the digital stories with a peer, engage in readers’ theatre, put on a puppet show, and so forth. *Extension activities* often included opportunities for the students to engage in collaborative work. For example, after reading *The Fruit Family* story, a kindergarten teacher could have her students draw pictures of the different types of fruit that they ate and label their pictures.

Teachers were informed that while regular access to appropriately leveled and progressively more demanding word-level and text-level activities were required, this suggested curriculum was a

fairly flexible guideline and should be adapted to meet the individual needs of their students as well as their own teaching styles. Teachers also had freedom to conduct the intervention as whole class, small group, individual, or some combination of the groupings. It was known at training that teaching would also vary depending upon access to technology in particular schools (e.g., presence of interactive whiteboard systems known as SMART Boards [SMART Technologies, Calgary, Alberta, Canada] or multiple computers in classrooms, use of a distinct ICT rooms in school). The responsibility for developing appropriate specific lesson plans and interventions always rested with the regular classroom teachers.

All of the teachers received a hard copy of an extensive ABRA teacher’s manual that illustrated in detail how ABRA could be used in these domains (Abrami, White, & Wade, 2010). Teachers also visited the Teacher’s Zone (available online at <http://abralite.concordia.ca/pd/index.php>), a resource area for them. Finally, the teachers met in small groups according to the grade levels they taught and planned ABRA introductory lessons for their classes based on the suggested 1-hr format.

Project management and roles. In each of the three provincial sites, an advanced graduate student coordinated the execution of the study under the supervision and support of a project literacy coordinator and a lead researcher in each province. Each provincial site’s co-coordinator was then responsible of recruiting and overseeing the training of their team of research assistants (RAs). The RAs visited participating classrooms and administered the pre- and posttest measures. During the first 4 weeks of the implementation phase, an RA was present in every experimental classroom to provide technical support and answer general pedagogical questions regarding the utilization of ABRA. Most teachers requested support for the first 3–5 lessons. The RAs also acted as the liaison between the teachers and the rest of the ABRA team. Monthly team meetings were held, and regular correspondence took place to update and address questions about data collection and lesson support. In the fifth week of intervention, the RAs started collecting treatment integrity data.

Literacy Assessment Measures

As ABRACADABRA was designed to aid children in learning alphabets, phonics, phonological awareness, reading fluency, oral language comprehension, and reading comprehension, reliable and valid psychometric tests were selected to examine all of these component abilities.

Letter-sound knowledge. To assess letter-sound knowledge, an experimenter showed a participant the 26 letters of the English alphabet and asked the student to say the corresponding sound of each letter presented following the assessment and scoring system described by Savage et al. (2009). The Spearman–Brown split-half internal reliability (*r*) of this test in the present sample at pretest was .87.

Blending words. This measure assesses a child’s phonological blending ability. A subtest of the Comprehensive Test of Phonological Processing (CTOPP) was used to examine students’ ability to blend words (Wagner, Torgesen, & Rashotte, 1999). In this test, the children listen to a series of disjointed sounds and then blend the sounds together to make a whole word. The Spearman–

Brown split-half reliability coefficient for this measure in the present sample at pretest was .86.

Fry words. To assess the students' word reading skills, we adapted a test using words from the Fry's Instant Word List (Fry, Kress, & Fountoukidis, 2000). Twenty words were randomly selected from Fry's first 200 words. The same 20 words were used at pre- and posttest. Each of the selected 20 words was placed on individual index cards and shown one at a time to participants. The students read each word presented to them and received a point for each word correctly read. The maximum score for this test was 20. The Spearman-Brown split-half reliability of this test in the present sample at pretest was .89.

Group Reading Assessment and Diagnostic Evaluation (GRADE). The GRADE is a standardized, nationally normed instrument designed to be administered to either the whole class or individually (Williams, 2001). The GRADE is reported to have strong internal consistency (r s ranging from .95 to .99) and retest reliability ($r = .80$; Williams, 2001). Reviews of the GRADE (Fugate, 2003; McBride, Ysseldecke, Milone, & Stickney, 2010; Waterman, 2003) have concluded that this tool is a reliable and valid measure of early reading ability.

Listening Comprehension subtest of the GRADE. We used the Listening Comprehension subtest to assess the students' understanding of spoken language. Children are read sentences and then asked to select a picture from four choices that best illustrates the meaning of each sentence. The Spearman-Brown split-half internal reliability coefficient for this measure in our sample at pretest was .68.

GRADE Sentence Comprehension and Passage Comprehension subtests. We used these Grade subtests to assess reading comprehension skills. The GRADE technical manual asserts that the reading comprehension test measures metacognitive skills of previewing, predicting, clarifying, and summarizing, making it a good measure of the comprehension skills taught in ABRA. The *Sentence Comprehension subtest* is a cloze procedure whereby students read a sentence that has a missing word and then select one of four words that best fits the context. In the *Passage Comprehension subtest*, students read short passages of around 32 words and then select the correct multiple-choice response to best answer each question. The Spearman-Brown split-half internal reliability coefficient in our sample at pretest was .76 for the sentence comprehension and .69 for passage comprehension.

Phonemic Segmentation Fluency subtest. The Phonemic Segmentation Fluency subtest of the Dynamic Indicators of Basic Literacy Skills (DIBELS; Kaminski & Good, 1996) assesses students' ability to fluently break three- or four-phoneme words into their individual phonemes in 1 min. For example, if the RA read *ship*, students had to say /sh/ /i/ /p/. The Spearman-Brown split-half reliability coefficient for this measure in the present sample at pretest was .97.

Oral Reading Fluency and Retell Fluency subtests. In the DIBELS, Oral Reading Fluency (ORF) subtest measures students' ability to read a passage out loud for 1 min. The number of correct words per minute from the passage yields the oral reading fluency rate. Students who correctly read 10 or more words in the DIBELS ORF task were then administered the Retell Fluency measure. Here, students tell the examiner all they can remember about the

passage in 1 min. The number of words retold yields the Reading Retell Fluency score.

Testing Procedure

All participants completed the Letter-Sound Knowledge, Fry Words, Blending Words, and GRADE Listening Comprehension measures. Other measures examining higher level literacy skills (reading fluency and reading comprehension) were developmentally appropriate only for students in Grades 1 and 1/2 at posttest. Fluency and comprehension measures were not administered to any of the kindergarten students at either pre- or posttest.

All children were seen twice at both pre- and posttest for testing. The first session involved individual testing of children. Here the Fry Words, Letter-Sound Knowledge and CTOPP Blending tasks were administered. Additionally, at posttest only for the Grade 1 and Grade 1/2 children, the three DIBELS subtests were also administered. In the second session, all GRADE measures were administered using a whole-class group-testing approach for children in two provincial sites and were individually administered in a third provincial site.

Treatment Integrity

In this study, a theoretically underpinned treatment integrity measure was developed that focused on the implementation of ABRA in the classrooms, following recommendations for best practice in assessing treatment integrity (Mowbray, Holter, Teague, & Bybee, 2003). Pairs of RAs first independently observed and recorded activities in around 20% of ABRA intervention classes. After each RA independently recorded his or her observations, they compared their ratings for agreement. Most pairs agreed on 98% of content of all observed sessions.

Observations were then formally coded using the Implementation Fidelity Measure (IFM), which was based on a rubric reported by Savage et al. (2010). Savage et al. adapted a classification system from Sandholtz, et al.'s (1997) stage model of computer technology integration by teachers that we described earlier, namely, entry, adoption, and adaptation. The criteria specified under each category of ABRA implementation also closely match the first three of Sandholtz et al.'s model criteria described earlier. Our first aim was to identify adequate treatment integrity. Teachers were informed during their training that adequate treatment integrity would consist of 20 hr of exposure to ABRA for each child in the intervention condition and that adequately implemented ABRA lessons would show evidence of careful planning, differentiation, and progression (e.g., in task difficulty) over time. Our aim, however, was to identify not just compliance with treatment but also possible excellence in implementation of ICT in classroom pedagogy. Whereas entry and adoption reflect traditional concerns about weak versus compliant intervention, adaptation refers to compliant intervention (as measured by reaching the same criteria for faithful *adoption* of ABRA) but *also* to clearly evident use of technology to transform wider learning in the classrooms. Two additions to the ABRA IFM rubric included the addition of a control level for teachers who did not use any aspect of ABRA technology, and a fifth category labeled as *differentiated adaptation* to highlight transformative use of ABRA in teaching across multiple skill levels including both word-level reading skills and

higher text-level comprehension and fluency extension activities. This latter category was distinguished from adaptation, which involved the transformative use of technology in *one* domain of literacy, most often in word-level reading skills. In understanding the differentiated adaptation measure, it should be noted that the use of ABRA at both word-level and higher text-level comprehension and fluency levels was encouraged but was never presented to teachers as a requirement of faithful adoption of ABRA. See Figure 2 for details of levels of implementation fidelity.

To establish rater reliability, we conducted a test trial of the rubric among four raters on four completed classroom observations. From a total of 12 independent IFM ratings, 11 of the judgments were identical across the four raters, thus establishing a 92% rater-reliability for the IFM rubric. Subsequently, every observation checklist completed during this study ($n = 293$) was rated by two of these raters. The raters had perfect agreement on 263 of the 293 IFM observation checklists they categorized. This is an 89% interrater agreement (Cohen's $\kappa = .83, p < .001$). The raters had 100% agreement in their ultimate categorization of teachers' use of technology on the 0–4 IFM scale.

Assessment of the general pedagogical quality of classrooms. To obtain further information about the quality of teaching, irrespective of whether teachers were implementing ABRA, observations of nontechnology based on regular language arts teaching was undertaken in both experimental and control classrooms. Senior RAs with previous early literacy experience conducted

observations using the Early Literacy and Language Classroom Observation (ELLCO; (Smith, Dickinson, & Sangeorge, 2002). The ELLCO is a standardized instrument assesses both the global quality of the classroom (e.g., classroom climate, approach to management, classroom organization), as well as language, literacy, and curriculum. The ELLCO takes about 90 min to complete. A brief teacher interview was also a part of this measure. In one provincial site ($n = 36$ classrooms), pairs of RAs observed the same lesson, rated the classes, and then met after the lesson to come to a rating consensus about the quality of classroom supports for literacy. Due to the availability conflicts of qualified RAs with an early literacy background, one rater, rather than two, collected data on this measure in two of the three provincial sites. Finally, we asked all teachers to complete a literacy intervention questionnaire (LIQ) devised by the current authors to ask about salient aspects of teachers' implementation of ABRA and teachers' attitudes toward the ABRA technology.

Results

Preliminary Data Analyses

Prior to analyses, data were first collated into separate child- and classroom-level spreadsheets. Preliminary data analyses suggested that there was no marked kurtosis or skewness in the classroom-level attainment data. All classroom-level variables were within

ABRA – IMPLEMENTATION RUBRIC

CONTROL (0)	ENTRY (1)	ADOPTION (2)	ADAPTATION (3)	DIFFERENTIATED ADAPTATION (4)
<ul style="list-style-type: none"> ◆ No implementation ◆ No aspect of ABRA was used 	<ul style="list-style-type: none"> ◆ Little to No evidence of teacher planning of ABRA ELA lessons ◆ Little to No evidence of teacher instructional guidance ◆ Little to No evidence of teacher monitoring students' use of ABRA ◆ No evidence of teacher's awareness of zone of proximal development as students instructed to all work on same activity level ◆ Minimal student exposure to ABRA activities ◆ ABRA exposure mainly unstructured lessons where students choose own activities (play-time / free-time) ◆ Occasional disruption and off-task behaviour of students ◆ Unclear about teacher and student's navigational comfort level with ABRA OR ◆ Teacher frustration/discomfort with technology evident 	<ul style="list-style-type: none"> ◆ Basic evidence of teacher planning of ABRA lessons ◆ Basic evidence of teacher instructional guidance of ABRA ◆ Basic evidence of teacher monitoring students' use of ABRA ◆ Some evidence of teacher's awareness of zone of proximal development as students are instructed to move up task levels if too easy / completed or move back if too hard ◆ ABRA exposure evidence of structured lessons ◆ Little off-task behaviour of students ◆ Teacher and students appear comfortable with navigating through ABRA activities ◆ Some evidence of differentiated use of ABRA activities, but mainly within one skill level (i.e. Phonics / Word Level activities). 	<ul style="list-style-type: none"> ◆ Clear evidence of teacher planning of ABRA lessons. Teacher links planning and target setting according to students ability level ◆ Clear evidence of teacher providing appropriate instructional guidance / feedback while students on ABRA ◆ Clear evidence of teacher monitoring students' use of ABRA ◆ Clear evidence of teacher's awareness of zone of proximal development as students are instructed to move up task levels if too easy / completed or move back if too hard ◆ ABRA exposure – evidence of structured lessons ◆ Students are clearly engaged in the lesson ◆ Teacher & students are comfortable with navigating through ABRA activities ◆ Extension of ABRA – Some evidence of entry-level activities that extend skills explored in one domain of ABRA, usually Word level (i.e., Rhyme matching game; spelling words or simple sentences; playing BINGO, etc). ◆ Evidence of differentiated use of ABRA activities, at more than one skill level (i.e. Word Level, Text Level or Writing activities). ◆ Evidence of collaborative work & use of collaborative learning opportunities 	<p>Criteria as per level four with the addition of the following:</p> <ul style="list-style-type: none"> ◆ Extension of ABRA – Clear evidence of extension activities that incorporate higher-level skills (i.e. Comprehension) that extend beyond simple WORD level activities. Examples such as, writing alternate story endings; Journal entry reflections on ABRA story; creating a drama skit/puppet show based on ABRA story, etc. ◆ Teacher clearly differentiated use of ABRA across ALL FOUR suggested levels of Implementation (i.e. Word Level, Text Level, Collaborative Work & Extension Activities) ◆ Teacher uses all collaborative learning opportunities – peer supported dialogues, different roles, reciprocal tutoring, etc.

Original Observer / Class / Date: _____

RATER: _____ DATE: _____

SCORE:

Figure 2. ABRA Implementation Fidelity Measure (IFM) rubric. ABRA = acronym for ABRACADABRA (A Balanced Reading Approach for Canadians Designed to Achieve Best Results for All); ELA = English language arts.

acceptable limits of normality for skewness and kurtosis, so no data transformations were undertaken. There was no strong evidence of outliers. At this point, no data were excluded from analysis. In the case of classroom-level data, there was some missing data. In eight kindergarten classrooms, Listening Comprehension subtests were not administered by error. Four teachers did not wish children to be assessed on the Reading Comprehension and Fluency measures. In each case, this data loss was at the school level so that paired intervention and control classes were lost, thus producing no unbalancing of the overall RCT design. Imputation of an entire classroom’s set of scores is not possible, so analyses for Listening Comprehension are based on 66 paired classrooms (74 – 8) and the Reading Comprehension and Fluency are based on 46 (50 – 4) paired classrooms.

Student-level missing data. We inspected the pattern of student-level missing data using the SPSS Missing Value Analysis (MVA) package to consider the randomness and impact of missing data and to then impute missing values where most appropriate. Missing data represented less than 5% of the total data across all variables, and so most common procedures for dealing with missing values would yield similar results (Tabachnick & Fidell, 2007). Analyses were performed using conventional chi-square tests to contrast the proportion of missing versus present data in intervention versus control group conditions at pre- and posttests. There was no evidence of selective experimental mortality across conditions, $\chi^2(1) < 1$, *ns*, in all cases. Further analysis using MVA revealed full-sample child-level data were also missing completely at random (MCAR) for all variables, according to Little’s MCAR test ($p > .05$ in all cases). Regression-based imputation procedures were selected, with reading pretest variables serving as predictors. The means of the fifth iteration were selected for analyses.

Main Data Analyses

The classroom-level means and standard deviations of all classroom-level attainment variables at pre- and posttest are pre-

sented in Table 1. Inspection of Table 1 shows signs of posttest advantage for the ABRA group on Letter-Sound Knowledge, Phonological Blending, Fry Words, and Listening Comprehension. Few clear signs of advantage were evident in the Grade 1 Reading Comprehension and Fluency measures.

Does a classroom-level cluster RCT effectiveness trial of ABRA yield significant advantages for intervention over control classrooms in early literacy at posttest? In this cluster RCT, random allocation of students took place at the classroom level, producing a nested design in which the likely contextual influence that classrooms may have on the achievement of the individual participants can be evaluated (e.g., Hox, 2010; Raudenbush & Bryk, 2002). Our data were thus first analyzed with hierarchical linear modeling (HLM) with randomized classroom as the unit of analysis. The final HLM models were built in standard bottom-up fashion from preliminary analyses with steps in HLM followed sequentially in order to yield the final models. Model 1 was an unconditional one-way analysis of variance (ANOVA) model with random effects and confirmed that there was classroom-level variance at pretest and posttest on attainment measures beyond variance attributable to pupils and that HLM was appropriate.

Subsequent ANCOVA models tested whether candidate covariates were significant and should be retained in the final model. An ANCOVA model was appropriate in this design as pretest attainment, and chronological age was always a significant covariate of its corresponding posttest measure. The final three-level hierarchical model, built on these tested assumptions, sought to establish whether the significant classroom-level variance on posttest attainment measures (after control for school-level shared variance at Level 3, pretest classroom-level attainment variance at Level 2, and pre- and posttest pupil level attainment variance and pupil chronological age at Level 1) was explained by the ELA with ABRA versus ELA without ABRA factor. Equations 1, 2, and 3

Table 1
Comparison of Raw Score Means, Standard Deviations, and Adjusted Means by Experimental Condition

Measure	Intervention groups					Control group				
	Pretest		Posttest			Pretest		Posttest		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	Adjusted ^a <i>M</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	Adjusted ^a <i>M</i>
Kindergarten & Grade 1/2										
Letter-Sound Knowledge	16.01	6.24	21.26	4.17	21.44	16.57	6.76	20.63	4.81	20.45
Fry Words	5.90	5.10	11.43	6.09	11.80	6.85	5.81	11.53	6.64	11.17
CTOPP Blending Words	6.41	2.94	9.96	3.40	10.27	7.19	3.04	9.48	3.56	9.17
DIBELS Phonemic Segmentation	—	—	28.75	11.33	29.78 ^b	—	—	27.60	11.41	26.57 ^b
GRADE Listening Comprehension	14.14	1.40	15.54	1.05	15.60	14.58	1.22	15.57	1.03	15.41
Grade 1/2										
DIBELS Reading Fluency	—	—	30.72	17.85	33.48 ^c	—	—	38.32	21.95	35.57 ^c
DIBELS Reading Retell	—	—	10.83	7.23	11.84 ^c	—	—	12.39	8.05	11.38 ^c
GRADE Sentence Comprehension	—	—	8.33	3.40	8.85 ^c	—	—	9.26	3.68	8.73 ^c
GRADE Passage Comprehension	—	—	9.00	3.44	9.57 ^c	—	—	9.94	3.95	9.38 ^c

Note. CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Literacy Skills; GRADE = Group Reading Assessment and Diagnostic Evaluation.
^a Adjusted means calculated using pretest and chronological age means as covariates. ^b Calculations for Phonemic Segmentation derived from pretest Blending Words mean. ^c Adjusted means for all Grades 1 & 2 reading measures (Reading Fluency, Reading Retell, Sentence Comprehension, & Passage Comprehension) calculated with the Fry Words pretest mean scores.

describe this final model at the pupil, classroom, and school levels, or student i in classroom j in school k , respectively.

Equation for Student Level-1 Model: (1)

$$\beta_{00k} = \gamma_{000} + \mu_{00k}$$

$$Y_{ijk} = \pi_{00k} + \pi_{1jk}(\text{PRETEST}_{ijk}) \\ + \pi_{2jk}(\text{CHRONOLOGICAL AGE}_{ijk}) + e_{ijk}$$

Equation for Classroom Level-2 Model: (2)

$$\pi_{00k} = \beta_{00k} + \beta_{01k} * (\text{PRETEST ATTAINMENT}_{jk}) \\ + \beta_{02k} * (\text{INTERVENTION}_{jk}) + r_{0jk}$$

Equation for School Level-3 Model: (3)

In these analyses, predictor variables were left uncentered and ratio-level raw scores which have a meaningful zero point value were used so as to ease interpretation. Equations 1–3 also show that the slope coefficients for all independent variables are treated as fixed: They are not allowed to randomly vary across classrooms and schools.

The results of these analyses are reported in Table 2. The measures administered to both kindergarten and Grades 1 and 2 are reported in Section A of Table 2. This analysis is of four measures—Letter-Sound Knowledge, Fry words, Phonological Blending, and Listening Comprehension—across all 74 paired classrooms (66 in the case of Listening Comprehension). These

Table 2
Hierarchical Linear Models Results for the Effect of ABRA Condition on Posttest Attainment

Variable	Student-level model		Classroom-level model	
	Coefficient	SE	Coefficient	SE
Section A ($n = 74$) Kindergarten & Grade 1/2 classrooms				
CTOPP Phonological Blending 2 = Dependent variable				
Chronological age	0.043	0.014**		
CTOPP Phonological Blending 1			0.631	0.024**
Intervention condition			1.080	0.320**
Letter-Sound Knowledge 2 = Dependent variable				
Chronological age	0.054	0.019**		
Letter-Sound Knowledge 1			0.518	0.018**
Intervention condition			0.961	0.461*
Fry Words 2 = Dependent variable				
Chronological age	0.099	0.021**		
Fry Words 1			0.629	0.022**
Intervention condition			0.691	0.690
DIBELS Phonemic Segmentation 2 = Dependent variable				
Chronological age	−0.308	0.062		
DIBELS Phonological Blending 1			1.508	0.109**
Intervention condition			2.578	1.425†
GRADE Listening Comprehension 2 = Dependent variable				
Chronological age	−0.007	0.008		
GRADE Listening Comprehension 1			0.365	0.023**
Intervention condition			0.085	0.160
Section B ($n = 50$) Grade 1/2 classrooms				
DIBELS Reading Fluency 2 = Dependent variable				
Chronological age	−0.177	0.153		
Fry Words 1			3.658	0.133**
Intervention condition			−2.517	2.282
DIBELS Reading Retell 2 = Dependent variable				
Chronological age	.119	.070*		
Fry Words 1			1.032	0.062**
Intervention condition			−0.148	0.969
Section B ($n = 46$) Grade 1/2 classrooms				
GRADE Sentence Comprehension 2 = Dependent variable				
Chronological age	−0.032	0.028		
Fry Words 1			0.584	0.025**
Intervention condition			0.148	0.375
GRADE Passage Comprehension 2 = Dependent variable				
Chronological age	0.012	0.031		
Fry Words 1			0.512	0.027**
Intervention condition			−0.027	0.440

Note. ABRA = shortened version of the acronym ABRACADABRA (A Balanced Reading Approach for Canadians Designed to Achieve Best Results for All) software program; CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Literacy Skills; GRADE = Group Reading Assessment and Diagnostic Evaluation.

† $p < .1$. * $p < .05$. ** $p < .01$.

results show that there is a significant effect of ABRA on intervention mean versus control group mean contrasts for Phonological Blending (10.27 vs. 9.17, $\beta = 1.08$, $p < .01$) and Letter-Sound Knowledge (21.44 vs. 20.45, $\beta = 0.96$, $p < .05$), and marginally for Phoneme Segmentation Fluency (29.78 vs. 26.57, $\beta = 2.58$, $p < .1$). Other effects however did not reach significance. Turning to the six measures of reading comprehension and fluency administered only to children in Grades 1 and 2 (50 paired classrooms), results showed no overall significant effects of ABRA on these measures.

Do high-implementing classrooms yield additional effects for ABRA? Prior to analyses, inspection of the IFM data revealed that of the 37 ABRA intervention teachers, there were seven entry-level teachers, 20 adoption-level teachers, seven adaptation-level teachers, and three differentiated-adaptation-level teachers. Due to the very small size of the last Time 1 (T1) category, the adaptation-level and differentiated-adaptation-level teachers were combined into a single adaptation-level category with 10 teachers. Of the entry-level teachers, six were kindergarten teachers, and one was a Grade 1 teacher. The means, standard deviations, and effect sizes by T1 condition are reported in Table 3 with the analyses involving the whole sample reported in Section A and those involving the Grade 1/2 in Section B of the table. There was evidence of value-added effect sizes for Phonological Blending, Letter-Sound Knowledge, and Phoneme Segmentation Fluency in the adoption and adaptation conditions relative to the control and entry conditions.

Our research questions here were specific and directional: We predicted largest improvements in literacy measures in adoption classrooms and high-implementing adaptation ABRACA-DABRA classrooms relative to control classrooms. Three-level (school, classroom, and pupil) HLM covariance models parallel to those used in the main analysis described earlier were used here. The only difference was that in these analyses, treatment integrity was considered as a factor entered to explore the relative differences in reading-dependent variable measures, with the control group as the reference category:

Equation for Student Level-1 Model: (4)

$$Y_{ijk} = \pi_{00k} + e_{ijk}$$

Equation for Classroom Level-2 Model: (5)

$$\pi_{00k} = \beta_{00k} + \beta_{01k} * (\text{ENTRY}_{jk}) + \beta_{02k} * (\text{ADOPTION}_{jk}) \\ + \beta_{03k} * (\text{ADAPTATION}_{jk}) + r_{0jk}$$

Equation for School Level-3 Model: (6)

$$\beta_{00k} = \gamma_{000} + \mu_{00k}$$

Each Level-2 predictor was treated as fixed and not allowed to randomly vary across schools.

As Table 3 shows, the predicted scores for each of the three groups (entry vs. control, adaptive vs. control, and adoptive vs. control) were then used in subsequent planned comparisons. These analyses revealed that the adoption group had significantly higher mean performance than the control group in Phonological Blend-

ing (10.52 vs. 9.48, $z = 2.55$, $p = .01$), Phoneme Segmentation Fluency (31.81 vs. 26.94, $z = 3.07$, $p < .01$), Sight Word Reading (12.87 vs. 11.26, $z = 1.99$, $p < .05$), and marginally for Letter-Sound Knowledge (21.46 vs. 20.69, $z = 1.41$, $p = .08$). The adaptation group similarly showed significantly higher performance than the control group on Letter-Sound Knowledge (22.58 vs. 20.69, $z = 2.37$, $p < .01$), Phonological Blending (10.70 vs. 9.48, $z = 1.97$, $p < 0.01$), and Phoneme Segmentation Fluency (31.30 vs. 26.94, $z = 1.85$, $p < .05$). No advantages were evident in any analysis for the entry group. This analysis confirms the presence of comparable significant advantages over controls for both the adoption and adaptation teachers who together made up 80% of the intervention teacher sample.

A final issue in the experimental study was the comparison of 74 intervention and control classrooms on background ELLCO measure four subscales—Language Literacy Curriculum, Classroom Observation, Book Reading, and Writing—and combined total score—Literacy Activities Overall Rating Scale total. The question this analysis addressed was whether ABRA intervention and control classrooms are comparable on the observed quality of their non-ABRA regular ELA teaching. The mean scores and results of these analyses are shown in Table 4. ANOVA analyses show that intervention and control group teachers are highly comparable both on overall ELLCO and on subscale measures, all $F_s < 1$, ns , except for the Writing subtest, where $F(1, 72) = 2.67$, $p = .11$. Analysis was undertaken on ELLCO scores across T1 category to explore whether the stronger implementers of ABRA were simply better teachers overall than the less-strong implementers. The results are reported in Table 5. These ANOVA analyses show that there were no significant differences in ELLCO score by implementation type, all $F_s < 1.45$, ns , except for the Writing subtest, where $F(2, 64) = 2.86$, $p = .07$, consistent with the view that the technology-implementers were not simply better literacy teachers overall.

Explorations of poor implementation fidelity. Beyond the experimental contrasts previously described, an important issue in any educational field study concerns understanding why teachers chose either not to implement the prescribed intervention or to implement it suboptimally. The issue is important because relevant evidence can guide future attempts to improve implementation and provides a sense of whether take-up of interventions is achievable. The present article carefully explored treatment integrity, focused on extended training of teachers, observed teachers' practice, and asked teachers about their practice through questionnaires and discussions. Detailed records of all of these elements were recorded to assess whether scale-up of the study was achievable and under what conditions.

What is known from this data of patterns of treatment integrity? First our direct observations reported in Table 3 showed that of 37 teachers asked to implement ABRA in this study, only seven were judged to have performed at the entry level of poor implementation. All 30 other teachers (80% of the intervention sample of teachers) achieved at least an adoption level of implementation or adopted the program faithfully as well as adapting the teaching in their classrooms, and as described earlier, both of these groups were associated with significant gains in a range of early literacy measures. As noted earlier, of the seven entry-level teachers, only one was a Grade 1 teacher, the other six all being kindergarten teachers. Thus, there was 96% adoption of ABRA among the

Table 3
Predicted Means, Standard Errors, and Adjusted Means by Treatment Quality Groups: Kindergarten and Grade 1/2 Measures

Measure/Treatment quality level	Pretest			Posttest			Adjusted <i>M</i>
	<i>n</i>	Mean	<i>SE</i>	<i>n</i>	Mean	<i>SE</i>	
Letter-Sound Knowledge							
Control	37	16.50	1.07	37	20.64	0.74	20.69 ^a
Entry	7	10.25	2.29	7	17.31	1.07	21.03 ^a
Adoption	20	18.64	1.49	20	22.82	2.43	21.46 ^a
Adaptation	10	14.57	2.17	10	21.33	4.30	22.58 ^a
Fry Words							
Control	37	6.91	0.99	37	11.53	6.64	11.26 ^a
Entry	7	2.49	1.89	7	6.29	2.26	9.85 ^a
Adoption	20	7.96	1.22	20	13.88	1.47	12.87 ^a
Adaptation	10	5.03	1.81	10	11.16	2.16	11.94 ^a
CTOPP Blending Words							
Control	37	7.18	0.55	37	9.70	0.67	9.48 ^a
Entry	7	4.14	1.03	7	7.77	1.16	9.82 ^a
Adoption	20	7.71	0.66	20	11.17	0.74	10.52 ^a
Adaptation	10	6.17	0.99	10	10.19	1.11	10.70 ^a
DIBELS Phonemic Segmentation							
Control	—	—	—	37	27.61	1.93	26.94 ^b
Entry	—	—	—	7	16.84	3.88	23.41 ^b
Adoption	—	—	—	20	33.63	2.51	31.81 ^b
Adaptation	—	—	—	10	30.84	3.72	31.30 ^b
GRADE Listening Comprehension							
Control	33	14.54	0.25	33	15.58	0.20	15.51 ^c
Entry	6	14.15	0.49	6	15.74	0.41	15.83 ^c
Adoption	19	13.91	0.28	19	15.41	0.24	15.59 ^c
Adaptation	9	13.71	0.44	9	15.12	0.37	15.33 ^c
DIBELS Reading Fluency							
Control	—	—	—	25	38.95	4.33	29.55 ^c
Entry	—	—	—	6	8.15	14.24	34.96 ^c
Adoption	—	—	—	16	31.84	4.51	23.75 ^c
Adaptation	—	—	—	8	25.84	6.47	26.87 ^c
DIBELS Reading Retell							
Control	—	—	—	25	12.91	1.71	8.91 ^c
Entry	—	—	—	—	5.02	5.70	11.25 ^c
Adoption	—	—	—	16	11.11	1.80	7.88 ^c
Adaptation	—	—	—	8	10.65	2.58	9.82 ^c
GRADE Sentence Comprehension							
Control	—	—	—	23	9.28	0.81	7.76 ^c
Entry	—	—	—	—	5.85	2.09	9.48 ^c
Adoption	—	—	—	14	8.90	0.72	7.89 ^c
Adaptation	—	—	—	8	6.61	0.81	6.77 ^c
GRADE Passage Comprehension							
Control	—	—	—	23	10.28	0.82	8.61 ^c
Entry	—	—	—	6	4.51	2.17	7.61 ^c
Adoption	—	—	—	14	10.44	0.74	9.36 ^c
Adaptation	—	—	—	8	7.00	1.00	6.85 ^c

Note. The covariate measures used in calculating the adjusted mean values for the reading measures were the pretest Fry Words and chronological age of students at pretest. CTOPP = Comprehensive Test of Phonological Processing; DIBELS = Dynamic Indicators of Basic Literacy Skills; GRADE = Group Reading Assessment and Diagnostic Evaluation.

^a Adjusted means calculated using corresponding pretest and chronological age as covariates. ^b Adjusted means for DIBELS Phonemic Segmentation calculated using pretest CTOPP Blending Words and chronological age value of students as covariates. ^c Adjusted means calculations for all Grades 1 and 2 reading measures (Fluency, Retell, Sentence Comprehension, & Passage Comprehension) used the Fry Words pretest scores.

Grade 1 or Grade 1/2 teachers, but only 50% adoption among kindergarten teachers in our sample, a by-grade difference that was highly significant, $\chi^2(1) = 14.30, p = .0002$.

On the other hand, it was very rare to find teachers in any grade in this study who adapted their teaching with ABRA to transform wider classroom teaching in both of the domains of word-level analysis and text-level comprehension. Indeed the very small numbers here (three from 37 or less than 10% of implementing teachers) meant that a combined measure of

adapted teaching was used to describe any ABRA teaching that transformed any aspect of classroom literacy instruction. Of the three teachers who did use such a transformative approach, it may have been that they had greater expertise in both literacy and technology to allow this to happen. In at least one case of the three, for example, the teacher had recently returned to the classroom after operating as a school board literacy consultant and was highly familiar with white board technologies in education.

Table 4
One-Way Analysis of Variance for Differences Between ABRA Condition Groups as a Function of Early Literacy and Language Classroom Observation (ELLCO) Measure Subtotals

ELLCO measure subtotals	Intervention group (n = 37)		Control group (n = 37)		ANOVA differences between ABRA condition groups		
	M	SD	M	SD	df	F	p
Language literacy curriculum	31.81	4.67	31.51	5.12	1	0.07	.80
Classroom observation	55.22	7.99	54.68	8.73	1	0.08	.78
Book reading	5.22	1.80	5.16	1.63	1	0.02	.89
Writing	3.76	.98	3.38	1.01	1	2.67	.11
Literacy activities overall rating scale total	8.97	1.92	8.54	1.88	1	0.96	.33

Note. ANOVA = analysis of variance; ABRA = shortened version of the acronym ABRACADABRA (A Balanced Reading Approach for Canadians Designed to Achieve Best Results for All) software program.

Analyses of the ELLCO scores (performed with one-sample *t* tests) of the one entry-level Grade 1 teacher against the 25 other intervention Grade 1 teachers showed that this teacher achieved significantly lower ELLCO scores on four of the five ELLCO measures than their peers, *t*(25) = 3.36, *p* = .003 for Language, Literacy, and Curriculum; *t*(25) = 4.57, *p* < .001 for Classroom Observation total; *t*(25) = 4.05, *p* < .001 for Book Reading; *t*(25) = -1.37, *ns*, for Writing, and *t*(25) = 2.97, *p* = .007 for Literacy Activities Rating Scale total. By contrast, the six entry-level kindergarten teachers were not so rated, compared with their kindergarten teachers peers in univariate ANOVAs contrasting them with the six adoption or adaptation teachers, *F*(1, 10) < 1, *ns*, for Language, Literacy, and Curriculum; *F*(1, 10) = 1.46, *ns*, for Classroom Observation total; *F*(1, 10) = 2.11, *ns*, for Book Read-

Table 5
One-Way Analysis of Variance for Differences Among ABRA Implementation Fidelity Measure Condition Groups as a Function of ELLCO Measure Subtotals

ELLCO measure subtotals/treatment integrity level	n	M	SD	ANOVA differences between treatment integrity groups		
				df	F	p
Language literacy curriculum				4	1.28	.29
Control	37	31.51	5.12			
Entry	7	29.14	5.90			
Adoption	20	32.80	4.36			
Adaptation	7	33.14	2.97			
Differentiated	3	31.66	4.87			
Classroom observation				4	1.43	.23
Control	37	54.68	8.73			
Entry	7	49.43	8.60			
Adoption	20	57.20	7.91			
Adaptation	7	57.00	4.55			
Differentiated	3	51.33	9.07			
Book reading				4	1.13	.35
Control	37	5.16	1.63			
Entry	7	4.00	1.63			
Adoption	20	5.50	1.57			
Adaptation	7	5.57	2.44			
Differentiated	3	5.33	1.53			
Writing				4	2.17	.08
Control	37	3.38	1.01			
Entry	7	4.14	.69			
Adoption	20	3.40	1.10			
Adaptation	7	4.14	.69			
Differentiated	3	4.33	.58			
Literacy activities overall rating scale total				4	.94	.44
Control	37	8.54	1.88			
Entry	7	8.14	1.35			
Adoption	20	8.90	1.97			
Adaptation	7	9.71	2.36			
Differentiated	3	9.67	1.53			

Note. ABRA = shortened version of the acronym ABRACADABRA (A Balanced Reading Approach for Canadians Designed to Achieve Best Results for All) software program; ELLCO = Early Literacy and Language Classroom Observation; ANOVA = analysis of variance.

ing; $F(1, 10) = 1.60$, *ns*, for Writing; and $F(1, 10) < 1$, *ns*, for Literacy Activities Rating Scale total. The six kindergarten entry-level teachers were also equally distributed across the three provinces of this study, and none were beginner teachers.

We inspected data from our LIQ teacher questionnaire. Data were available for three of the six entry-level kindergarten teachers, the one Grade 1 entry-level teacher, and four of six adoption- or adaptation-level kindergarten teachers. Responses to questions concerning the effects of technological difficulties did not show that entry-level teachers were more likely to experience technical difficulties implementing ABRA. Overall, half of both the entry- and adoption-level teachers commented on this issue as a factor. Teachers did not report that their knowledge of software was negatively reflected in reduced use of ABRA. Indeed, none of the entry-level teachers suggested this was an issue, whereas two of the four adoption- or adaptation-level teachers mentioned it as being an issue on occasion. These data are also consistent with a detailed analysis of the training of a subset of all teachers using ABRA in one site in this study reported by Anderson et al. (2011). Results of this study suggested dramatic reductions in requests for just-in-time assistance over the first 2 weeks or so of implementation of ABRA such that all teachers showed that they were well trained to use the technology competently and independently.

Finally, inspection of field notes from teacher training and discussions with staff leading the support for teachers were consistent with this general picture. Entry-level teachers were seen as being at least capable, and often excellent, teachers, but sometimes rather wary of technology or the use of any technology to teach reading in kindergarten on general pedagogical-philosophical grounds. At least some of these teachers drew attention to their provincial kindergarten curricula that did not encourage explicit phonological analysis of words or the teaching of reading at this early stage of children's careers. In sum, the majority of teachers (80% of intervention teachers) used ABRA with fidelity, and nearly all (96%) of Grade 1 teachers did so. Of the 20% who did not do so, the clearest feature of the group was that they were more likely to be kindergarten teachers, and while strong teachers, they sometimes articulated that on general pedagogical or curricular grounds that technology was not suitable or that the skills in ABRA were not suitable for children in kindergarten. On the other hand, even in kindergarten an equally large number of teachers used ABRA with fidelity, and as our analyses show, all such teaching showed significant advantages on a range of literacy measures.

Discussion

The main aim of the present study was to undertake a well-designed cluster RCT trial to evaluate the added value of ABRA delivered by regular classroom teachers in their regular kindergarten through Grade 2 ELA classes as a test of external validity of the ABRA resource. We sought to answer two questions: the first question concerned the overall effects of ABRA in intervention versus control group contrasts, and the second question concerned the relative effects of different levels of treatment integrity on outcomes.

Did ABRA Yield Significant Advantages for Intervention Over Control Classrooms in Early Literacy at Posttest?

Our analyses found that ABRA produced significant effects on standardized measures of Phonological Blending at posttest in 37 classrooms representing different school boards, regions, and educational policies from across Canada compared with 37 comparison ELA classrooms nested within the same schools that did not use ABRA between pre- and posttest. There were also significant effects of ABRA on Letter-Sound Knowledge and marginally significant effects on speeded phoneme segmentation ability. This specific pattern of effects of ABRA on phonological abilities replicates patterns reported in four previous studies of ABRA (Comaskey et al., 2009; Savage et al., 2009; Wolgemuth et al., 2011, 2012). The results of the present study are an important novel contribution to knowledge as this is the first time that an *effectiveness* study, that is, a study conducted by regular classroom teachers rather than an *efficacy* study conducted by research students or specially employed and trained educators, has shown ABRA to improve these aspects of literacy. Against this finding, and contrary to hypotheses, there were no significant effects of ABRA intervention on measures of Sight Word Reading, listening and reading comprehension, and fluency.

Theoretically, the development of phonological blending skills is closely allied to the ability to “decode” novel words. This ability has been described as the sine qua non of reading acquisition in alphabetic orthographies (e.g., Share, 1995, 2008). A large amount of evidence from experimental, longitudinal, and intervention research shows that children with strong blending and decoding skills have a greater likelihood of going on to be fluent readers and comprehenders of text than those with weak skills in these domains (e.g., Byrne, 1998; National Reading Panel, 2000; Savage, Carless, & Ferrero, 2007; Wagner et al., 1999). For these reasons, letter-sound knowledge, phonological blending skills, and decoding have appropriately been called foundations of literacy (e.g., Byrne, 1998; Seymour, 1997).

There were no significant effects for two other measures administered across the sample: Sight Word Reading and Listening Comprehension. There were also no significant effects for ABRA across five measures of comprehension and fluency, most of which were administered to children in Grade 1. It might be that the significant impact of ABRA on phonological awareness and letter knowledge at immediate posttest is reflected in significant growth in reading comprehension at a delayed posttest. Indeed, recent data suggest this pattern (Di Stasio et al., 2012). Nevertheless, the pattern here is somewhat disappointing, given that ABRA is designed to be a balanced literacy resource. There are a number of possible explanations for the patterns in the present study. One concern is that the smaller sample for these latter tests (48 classrooms) and the lower internal reliability of the comprehension tests may have impacted our ability to assess outcomes. However, the absence of trends favoring the intervention condition in sample estimates argues against the former interpretation. Alternatively, it is possible that ABRA does not implement comprehension and fluency strategies as well as it does phonic strategies. However, the finding that ABRA significantly improves comprehension and fluency in efficacy trials conducted by university researchers but not in teacher-led effectiveness trials argues against this view. A

more direct explanation of different patterns lies in the implementation of comprehension in teacher-led ABRA interventions. This issue of implementation is considered below.

Do Adequate and Transformational Implementations Yield Additional Effects for ABRA?

The second research question explored in the present study concerned the impacts of different levels of quality in implementation of ABRA. Detailed investigation of treatment integrity are rarely undertaken or reported fully (e.g., Campuzano et al., 2009; Lane et al., 2004). Detailed records were kept of 293 paired classroom observations of teachers using ABRA in the present study, yielding high interrater reliability for the coding system constructed. Further analysis of student data treating variation in the observed quality of teaching showed that ABRA effects were, as predicted, partly mediated by the quality of technology implementation by teachers. Teachers who were entry-level implementers with inadequate exposure to ABRA, poorer evidence of planning, differentiation, and progression in learning produced little growth in student literacy compared with the majority of teachers whose implementations were of superior quality at the adoption or adaptation levels of implementation. There was little evidence in size or extent of effects of an additional advantage for adaptation over adoption. This finding suggests that faithful adherence to delivery of the content of ABRA is the main source of the significant growth in literacy described here.

As noted earlier, there is much evidence that the quality of the implementation of programs has a significant impact on outcomes regardless of whether the interventions are successful and involve technology (Chambers, Abrami, et al., 2008; Savage et al., 2010; Wolgemuth et al., 2011; Wozney et al., 2006) or do not involve technology (e.g., Davidson et al., 2009; Lane et al., 2004; Stein et al., 2008). The current findings are consistent with this literature. However, it should be noted that unlike some studies (e.g., Chambers, Abrami, et al., 2008; Davidson et al., 2009), variation in program implementation did not *entirely* explain implementation effects. That is, overall significant effects of ABRA on phonic skills were evident across the whole sample irrespective of implementation quality. This is important as it shows that ABRA is effective across samples of teachers, given current abilities, while the treatment integrity effects suggest that if the largely kindergarten-based entry-level teachers can be supported to integrate or embed technology as skillfully as the adoption-level and adaptation-level teachers did, the impact of technology on literacy may be still greater still. Our analysis of these teachers' implementation suggests the need to win a wider debate about the suitability of educational technology in kindergarten and to be aware that local curricular constraints can influence uptake of technologies. In this sense, teacher professional development as well as the development of provincial curricula that fully support technology use may be required for maximal treatment integrity.

The same treatment integrity measure also showed that there were only three (two Grade 1/2 and one kindergarten classroom) teachers in the whole sample of 37 intervention teachers who adapted ABRA for both word-level decoding and word reading *and* text-level comprehension and fluency work. This finding was reported despite the fact that a substantial amount of time was spent covering these four important elements of literacy acquisi-

tion during the teacher training and subsequent feedback sessions. While our definition of treatment integrity called for teachers to use ABRA as they saw best for 20 hr per child in their classrooms, deliberate efforts were made during training sessions to have teachers focus on activities that addressed higher order skills. For example, training often began with a focus on comprehension and fluency sections and then later dealt with alphabets, so teachers would become familiar doing the "more difficult" comprehension activities first and then move on to the "easier" phonic skills. These views were then reiterated during the ongoing support and advice given to teachers. The finding that few teachers sought to use the comprehension components of ABRA provides both an obvious explanation of the null effects of ABRA on children's comprehension skill reported earlier and also suggests that further or more effective training must be done with teachers to incorporate the software within their classroom instruction.

In contextualizing these findings, it could be noted that there is relatively limited evidence that teachers use explicit strategy teaching for comprehension outside ICT (e.g., Pressley, 1998). It might be that in contrast to teaching phonics that, while not ubiquitous, is common in some form, teachers do not know how to explicitly teach cognitive strategies for comprehension. A second issue that should be borne in mind is that the ABRA intervention was conducted toward the beginning of the academic year with many beginner readers. It may be that teachers focus at this time on word reading and phonics, and balance is achieved in instruction not by concurrent equivalent emphasis (as required in this evaluation of ABRA) but rather by working more on comprehension later in the year. Certainly, recent observational studies in Grade 1 classes in some overlapping vicinities suggest that in the latter part of year, more of the focus is on comprehension and not exclusively on phonics (Deault, 2011; Deault & Savage, 2012), at least in Grade 1 classrooms.

Finally, it may be the case that the comprehension activities in ABRA do not go far enough in scaffolding the pedagogy of teachers and the learning of their students. Pressley (2002) concluded that there was little evidence of teachers encouraging students to become self-regulated readers by using good comprehension strategies. There was the expectation that students would be self-regulated in their reading, but no instruction of these self-regulated processes was given. Abrami, Lysenko, Wade, and Pillay (2011) linked the structured activities in ABRA with the self-regulation activities in ePEARL (LEARN, Laval, Quebec, Canada), an electronic portfolio tool. In a longitudinal quasi-experiment, Abrami et al. (2011) found positive effects on comprehension in medium- and high-implementation ABRA-ePEARL classrooms compared with control classrooms.

Turning to the wider literature on technology, as noted earlier, several comprehensive high-quality systematic reviews of RCT research suggest few positive effects of ICT on literacy (e.g., Andrews et al., 2007; Torgersen & Zhu, 2003). Much of the research included in these full historical reviews was undertaken in the 1980s and 1990s. More recently, high-quality RCT studies have been conducted around the world—for example, in France, (Ecalte et al., 2009), the United States (Chambers, Abrami, et al., 2008; Chambers, Cheung, Madden, Slavin, & Gifford, 2006; Chambers, Slavin, et al., 2008), Australia (with ABRA, Wolgemuth et al., 2011, 2012), and in Canada (with ABRA, Comaskey et al., 2009; Di Stasio et al., 2012; Savage et al., 2009). The current

ABRA research joins this small but growing literature. We thus concur with Andrews et al. (2007) who noted that it is a particularly exciting time for ICT research in literacy as both the theory of ICT and the evidence base drawn from this more sophisticated theoretical positioning is in its infancy. A more sophisticated theoretical position would also need to ask not just that technology works, but why it works.

Underlying Cause of the Effectiveness of ABRA

In a field-based effectiveness trial, the sources of influences on outcomes are, by definition, greater than in internal validity trials such as that reported by Savage et al. (2009). Effectiveness trials represent a distinct and crucial step in any evaluation of the utility of interventions. Why then did ABRA work here for phonic-related skills? Specifically, we have argued earlier that three overarching methodological issues are of the essence in high-quality ICT intervention: the *quality of the technology* used in research studies, the *quality of the implementation* of intervention studies, and the *theoretical and pedagogical coherence* of technologies. The quality of the ABRA technology, we argue, is high; for example, unlike most ICT, ABRA has a fully implemented synthetic phonics component (e.g., Grant et al., 2012). This may have supported the strong growth witnessed in phonic-related skills.

A second theme is that one needs to explore actions in the classroom in detail to understand the effects of implementation. Startlingly, some major intervention studies (e.g., Campuzano et al., 2009) do not explore implementation at all, so results tell us very little about an intervention. Another fact about implementation that might explain the effectiveness of ABRA is that ongoing and embedded support was provided to teachers. Unlike many training studies, here all teachers in this study participated in a training day and also received continuous support when delivering ABRA in their language arts classes. We gave constant support through regular visits and phone and e-mail advice in addition to detailed guidance on pacing the curriculum and feedback on teaching for the first 3–4 weeks. After that initial guidance, nearly all teachers no longer needed direct facilitator support to use ABRA. Our perception was that this initial direct support was crucial to the effective implementation of ABRA. Finally, we note that the Slavin, Lake, Chambers, Cheung, and Davis' (2009) systematic review of interventions suggested that the largest improvements found across intervention studies involve clear changes in teacher practice rather than changes to curriculum or other features of classroom life such as the use of ICT per se. For this reason alone, explorations of changes in teacher practice during and after interventions are of high importance.

The third issue concerns the theoretical and pedagogical coherence of technologies for literacy. We have been clear to articulate an embedded role for technology alongside of and highly connected to effective teaching in regular classrooms. Here technology can add value as an additional resource compared with effective teaching in the absence of technology. This model is of course reflected in our statistical analyses, where the value added by ABRA to classrooms was the focus. For this reason, the clinical trial model, while important, cannot be too rigidly applied as the best teaching with ICT is likely to be highly interconnected with other forms of learning. We have also pointed to a range of evidence sources that suggests that an implementation science of

reading needs to focus on what teachers do and also why they do and do not do things, and on the cultural aspects of the profession as much on interventions themselves. In ABRA specifically, there might thus be a role for exploring through “thick description,” the technology-based and nontechnology-based aspects of their classroom teaching before and well after ABRA has been introduced into classrooms.

Limitations of the Present Study

A number of potentially important potential limitations need to be noted in the present study. One issue concerns the relatively wide range in chronological age. There were, by chance, more children in Grade 2 in the control condition than in the intervention condition. Stratified randomization was not possible in small schools with one Grade 1 class and one Grade 1/2 class or larger ones with uneven numbers of classes at each grade level. It was judged to be methodologically superior to randomize within schools to control for the well-documented school-level effects on attainment, rather than to stratify across schools for age. The effects of these differences in age were probably modest: there was no significant difference between the intervention and control groups across the sample in chronological age, and effects reported persisted after control for chronological age. The results are potentially limited by reporting effects at immediate posttest and not also at delayed posttest. However, a precondition for obtaining the capacity to randomize interventions in schools was our promise to train the control class teachers after the intervention period was completed. As noted earlier, much longitudinal research, including our own work with ABRA, strongly suggests that the phonological and letter skills that improved here are foundational for later literacy. The patterns reported at posttest (e.g., significant effects for phonological and letter knowledge tasks overall and enhanced effects for high implementers) were also very closely related to observed teacher behaviors with ICT showing that gains observed are unlikely to be solely of a general motivational nature.

Implications and Future Studies

Arguably, the single biggest challenge facing reading researchers, implementation scientists, and practitioners in the 21st century is the issue of building scalable and sustainable interventions (Abrami et al., 2008). We demonstrated here that ABRA, an open, free-access resource, can be used effectively at scale across Canada. It has also used efficaciously in remote regions of northern Australia (Wolgemuth et al., 2011, 2012). ABRA has also been embedded within the practice of several school boards, ABRA's professional support staff, and others around the world. ABRA is thus a community resource in the long-established spirit of giving away the best and most useful findings of scientific psychology (Miller, 1969).

The present research also shows that such studies are possible for university-based teams to execute successfully in Canada. The culture of funding, collaborating between and in schools, and executing large-scale RCT studies is not at all widespread in Canada (e.g., Jamieson, 2006). Indeed, this study is the largest non-U.S. study of ICT to date. Indeed, as far as we are aware, this is largest true experiment (RCT) reading intervention of any sort reported to date in Canada. We would argue that this sort of work

yields clear value-added dividends for the community. The next step must be to explore the scale-up of interventions, to obtain a richer picture of the support that teachers need to encourage both high-level adaptations and strong expectancies of success, and to determine the long-term effects of intervention, ways to use evidence to design better and more effective ABRA activities, and how to best hand-over ABRA effectively to school boards. All of this is work in progress.

References

- Abrami, P. C., Lysenko, L., Wade, A., & Pillay, V. (2011). *A quasi-experimental study of the use of literacy and portfolios tools to support the comprehension skills of emerging readers*. Montreal, Quebec, Canada: Concordia University Centre for the Study of Learning and Performance.
- Abrami, P. C., Savage, R. S., Deleveaux, G., Wade, A., Meyer, E., & Lebel, C. (2010). The Learning Toolkit: The design, development, testing and dissemination of evidence-based educational software. In P. Zemliansky & D. M. Wilcox (Eds.), *Design and implementation of educational games: Theoretical and practical perspectives* (pp. 168–188). Hershey, PA: IGI Global. doi:10.4018/978-1-61520-781-7.ch012
- Abrami, P. C., Savage, R. S., Wade, C. A., Hipps, G., & Lopez, M. (2008). Using technology to assist children learning to read and write. In T. Willoughby & E. Wood (Eds.), *Children's learning in a digital world* (pp. 129–171). Oxford, England: Blackwell. doi:10.1002/9780470696682.ch6
- Abrami, P. C., White, B., & Wade, A. (2010). *ABRACADABRA LTK teacher guide*. Retrieved from http://grover.concordia.ca/abracadabra/resources/download/LTK_ABRA_lr.pdf
- Anderson, A., Wood, E., Piquette-Tomei, N., Savage, R. S., & Mueller, J. (2011). Evaluating the impacts of just-in-time instructional support for teachers introducing a web-based reading program for primary grade children. *Teaching and Teacher Education*, 19, 499–525.
- Andrews, R., Freeman, A., Hou, D., McGuinn, N., Robinson, A., & Zhu, J. (2007). The effectiveness of information and communication technology on the learning of written English for 5- to 16-year-olds. *British Journal of Educational Technology*, 38, 325–336. doi:10.1111/j.1467-8535.2006.00628.x
- Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Erlbaum.
- Bishop, D. V. M., & Snowling, M. J. (2004). Developmental dyslexia and specific language impairment: Same or different? *Psychological Bulletin*, 130, 858–886. doi:10.1037/0033-2909.130.6.858
- Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research*, 72, 101–130. doi:10.3102/00346543072001101
- Borenstein, M., Rothstein, H., Cohen, J., Schoenfeld, D., & Berlin, J. (2001). Power and precision (Version 2.1) [Computer software]. Englewood, NJ: Biostat.
- Byrne, B. (1998). *The foundation of literacy: The child's acquisition of the alphabetic principle*. Hove, England: Psychology Press.
- Campuzano, L., Dynarski, M., Agodini, R., Rall, K., & Pendleton, A. (2009). Effectiveness of reading and mathematics software products: Findings from two student cohorts (NCEE 2009–4041). Jessup, MD: National Center for Education Evaluation and Regional Assistance. Retrieved from <http://ies.ed.gov/ncee/pubs/>
- Chambers, B., Abrami, P., Tucker, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2008a). Computer-assisted tutoring in Success for All: Reading outcomes for first graders. *Journal of Research on Educational Effectiveness*, 1, 120–137. doi:10.1080/19345740801941357
- Chambers, B., Cheung, A. C. K., Madden, N. A., Slavin, R. E., & Gifford, R. (2006). Achievement effects of embedded multimedia in a success for all reading program. *Journal of Educational Psychology*, 98, 232–237. doi:10.1037/0022-0663.98.1.232
- Chambers, B., Slavin, R. E., Madden, N. A., Abrami, P. C., Tucker, B. J., Cheung, A., & Gifford, R. (2008b). Technology infusion in Success for All: Reading outcomes for first graders. *The Elementary School Journal*, 109, 1–15. Retrieved from http://successforall.org/_images/pdfs/Technology_Infusion_11_04_05.doc. doi:10.1086/592364
- Chen, C. H. (2008). Why do teachers not practice what they believe regarding technology integration? *The Journal of Educational Research*, 102, 65–75. doi:10.3200/JOER.102.1.65-75
- Chen, J. Q., & Chang, C. (2006). Using computers in early childhood classrooms. *Journal of Early Childhood Research*, 4, 169–188. doi:10.1177/1476718X06063535
- Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3, 149–210. doi:10.1007/BF01320076
- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, 18, 947–967. doi:10.1016/S0742-051X(02)00053-7
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, 100, 589–608. doi:10.1037/0033-295X.100.4.589
- Comaskey, E. M., Savage, R. S., & Abrami, P. C. (2009). A randomised efficacy study of Web-based synthetic and analytic programmes among disadvantaged urban kindergarten children. *Journal of Research in Reading*, 32, 92–108. doi:10.1111/j.1467-9817.2008.01383.x
- Cuban, L. (2001). *Oversold and underused: Computers in the classroom*. Cambridge, MA: Harvard University Press.
- Davidson, M. R., Fields, M. K., & Yang, J. (2009). A randomized trial study of a preschool literacy curriculum: The importance of implementation. *Journal of Research on Educational Effectiveness*, 2, 177–208. doi:10.1080/19345740902770028
- Deault, L. (2011). *Effective classroom contexts to develop literacy and attention skills for typical and at-risk first grade students* (Unpublished doctoral dissertation). McGill University, Montreal, Canada.
- Deault, L., & Savage, R. S. (2012). *Effective classroom contexts to develop literacy and attention skills for typical and at-risk first grade students*. Manuscript submitted for publication.
- Deault, L., Savage, R. S., & Abrami, P. C. (2009). Inattention and response to the ABRACADABRA Web-based literacy intervention. *Journal of Research on Educational Effectiveness*, 2, 250–286. doi:10.1080/19345740902979371
- Dede, C. (1996). Emerging technologies and distributed learning. *American Journal of Distance Education*, 10, 4–36. doi:10.1080/08923649609526919
- Di Stasio, M. R., Savage, R. S., & Abrami, P. C. (2012). A follow up study of the ABRACADABRA web based literacy intervention in Grade 1. *Journal of Research in Reading*, 35, 169–86. doi:10.1111/j.1467-9817.2010.01469.x
- Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., & Campuzano, L., . . . Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort*. Washington, DC: U. S. Department of Education, Institute of Education Sciences.
- Ecalle, J., Magnan, A., & Calmus, C. (2009). Lasting effects on literacy skills with a computer-assisted learning using syllabic units in low-progress readers. *Computers and Education*, 52, 554–561. doi:10.1016/j.compedu.2008.10.010
- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9, 167–188. doi:10.1207/s1532799xssr0902_4

- Ehri, L., Nunes, S., Willows, D., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, 36, 250–287. Retrieved from <http://www.jstor.org/stable/748111>. doi:10.1598/RRQ.36.3.2
- Fry, E. B., Kress, J. E., & Fountoukidis, D. L. (2000). *The reading teacher's book of lists* (4th ed.). Paramus, NJ: Prentice-Hall.
- Fugate, M. H. (2003). Review of the group reading assessment and diagnostic evaluation. In B. S. Plak, J. C. Impara, & R. A. Spies (Eds.), *The fifteenth mental measurements yearbook* (pp. 425–429). Lincoln, NE: Buros Institute of Mental Measurements.
- Fullan, M. (2001). *The new meaning of educational change* (3rd ed.). New York, NY: Teachers College Press.
- Grant, A., Wood, E., Gottardo, A., Evans, M. E., Phillips, L., & Savage, R. S. (2012). Assessing the content and quality of commercially available reading software programs: Do they have the fundamental structure to promote the development of early reading skills in children? *NHSA Dialog*, 15(4), 319–342. doi:10.1080/15240754.2012.725487
- Harasim, L., Hiltz, S. R., Teles, L., & Turoff, M. (1995). *Learning networks: A field guide to teaching and learning online*. Cambridge, MA: Massachusetts Institute of Technology Press.
- Hipps, G., Abrami, P. C., Savage, R. S., Cerna, N., & Jorgensen, A. (2005). ABRACADABRA: The research, design and development of Web-based early literacy software (pp. 89–112). In S. Pierre (Ed.), *DIVA: Innovations et tendances en technologies de formation et d'apprentissage* [DIVA: Innovations and trends in education and training technologies]. Montreal, Quebec, Canada: Presses Internationales Polytechnique.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- Hulme, C., Snowling, M., Caravolas, M., & Carroll, J. (2005). Phonological skills are (probably) one cause of success in learning to read: A comment on Castles and Coltheart. *Scientific Studies of Reading*, 9, 351–365. doi:10.1207/s1532799xssr0904_2
- Jamieson, D. G. (2006). Literacy in Canada. *Paediatrics & Child Health*, 11, 573–574.
- Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, 25, 215–227.
- Kay, R. H. (2008). Exploring the relationship between emotions and the acquisition of computer knowledge. *Computers & Education*, 50, 1269–1283. doi:10.1016/j.compedu.2006.12.002
- Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential—but often forgotten—component of school-based interventions. *Preventing School Failure*, 48, 36–43. doi:10.3200/PSFL.48.3.36-43
- Lewis, R. B., Ashton, T. M., Haapa, B., Kieley, C. L., & Fielden, C. (1999). Improving the writing skills of students with learning disabilities: Are word processors with spelling and grammar checkers useful? *Learning Disabilities: A Multidisciplinary Journal*, 9, 87–98.
- MacArthur, C. A., Ferretti, R. P., Okolo, C. M., & Cavalier, A. R. (2001). Technology applications for students with literacy problems: A critical review. *The Elementary School Journal*, 101, 273–301.
- Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge Press. doi:10.1017/CBO9781139164603
- McBride, J. R., Ysseldyke, J., Milone, M., & Stickney, E. (2010). Technical adequacy and cost benefit of four measures of early literacy. *Canadian Journal of School Psychology*, 25, 189–204. doi:10.1177/0829573510363796
- Miller, G. A. (1969). Psychology as a means of promoting human welfare. *American Psychologist*, 24, 1063–1075. doi:10.1037/h0028988
- Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *The American Journal of Evaluation*, 24, 315–340. doi:10.1177/109821400302400303
- Mueller, J., Wood, E., & Willoughby, T. (2008). The integration of computer technology in the classroom. In T. Willoughby & E. Wood (Eds.), *Children's learning in a digital world* (pp. 272–297). Oxford, England: Blackwell. doi:10.1002/9780470696682.ch11
- Mueller, J., Wood, E., Willoughby, T., Ross, C., & Specht, J. (2008). Identifying discriminating variables between teachers who fully integrate computers and teachers with limited integration. *Computers & Education*, 51, 1523–1537. doi:10.1016/j.compedu.2008.02.003
- National Reading Panel. (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups (National Institute of Health Publication No. 00-4769). Washington, DC: National Institute of Child Health and Human Development.
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11, 357–383. doi:10.1080/10888430701530730
- Pressley, M. (1998). *Reading instruction that works*. New York, NY: Guilford Press.
- Pressley, M. (2002). Metacognition and self-regulated comprehension. In A. Farstrup & J. Samuels (Eds.), *What research says about reading instruction* (3rd ed., pp. 291–309). Newark, DE: International Reading Association.
- Puffer, S., Torgerson, D. J., & Watson, J. (2005). Cluster randomized controlled trials. *Journal of Evaluation in Clinical Practice*, 11, 479–483. doi:10.1111/j.1365-2753.2005.00568.x
- Rabiner, D. L., & Malone, P. S. (2004). The impact of tutoring on early reading achievement for children with and without attention problems. *Journal of Abnormal Child Psychology*, 32, 273–284. doi:10.1023/B:JACP.0000026141.20174.17
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Sandholtz, J. H., Ringstaff, C., & Dwyer, D. C. (1997). *Teaching with technology: Creating student-centered classrooms*. New York, NY: Teachers College Press.
- Sang, G., Valcke, M., van Braak, J., & Tondeur, J. (2010). Student teachers' thinking processes and ICT integration: Predictors of prospective teaching behaviors with educational technology. *Computers & Education*, 54, 103–112. doi:10.1016/j.compedu.2009.07.010
- Savage, R. S. (2012). Evidence-based reading interventions: Implementation issues for the 21st century. In B. Kelly & D. F. Perkins (Eds.), *The Cambridge handbook of implementation science for psychology in education* (pp. 277–297). Cambridge, England: Cambridge University Press.
- Savage, R. S., Abrami, P. C., Hipps, G., & Deault, L. (2009). A randomized controlled trial study of the ABRACADABRA reading intervention program in Grade 1. *Journal of Educational Psychology*, 101, 590–604. doi:10.1037/a0014700
- Savage, R. S., Carless, S., & Ferraro, V. (2007). Predicting curriculum and test performance at age 11 years from pupil background, baseline skills, and phonological awareness at age 5 years. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 48, 732–739. doi:10.1111/j.1469-7610.2007.01746.x
- Savage, R. S., Deault, L., Daki, J., & Aouad, J. (2011). Orthographic analogies and early reading: Evidence from a multiple clue word paradigm. *Journal of Educational Psychology*, 103, 190–205. doi:10.1037/a0021621
- Savage, R. S., Erten, O., Abrami, P. C., Hipps, G., Comaskey, E., & van Lierop, D. (2010). ABRACADABRA in the hands of teachers: The effectiveness of a Web-based literacy intervention in Grade 1 language arts programs. *Computers & Education*, 55, 911–922. doi:10.1016/j.compedu.2010.04.002
- Savage, R. S., & Stuart, M. (2006). A developmental model of reading acquisition based upon early scaffolding errors and subsequent vowel

- inferences. *Educational Psychology*, 26, 33–53. doi:10.1080/01443410500340983
- Scardamalia, M., & Bereiter, C. (1996). Adaptation and understanding: A case for new cultures of schooling. In S. Vosniadou, E. DeCorte, R. Glaser, & H. Mandl (Eds.), *International perspectives on the design of technology-supported learning environments* (pp. 149–163). Mahwah, NJ: Erlbaum.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523–568. doi:10.1037/0033-295X.96.4.523
- Seymour, P. H. K. (1997). Foundations of orthographic development. In C. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to spell: Research, theory, and practice across languages* (pp. 319–337). London, England: Erlbaum.
- Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, 55, 151–218. doi:10.1016/0010-0277(94)00645-2
- Share, D. L. (2008). Orthographic learning, phonological recoding, and self-teaching. *Advances in Child Development and Behavior*, 36, 31–82. doi:10.1016/S0065-2407(08)00002-5
- Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, 43, 290–322. doi:10.1598/RRQ.43.3.4
- Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, 79, 1391–1466. doi:10.3102/0034654309341374
- Smith, M. W., Dickinson, D. K., & Sangeorge, A. (2002). *The early language and literacy classroom observation*. Baltimore, MD: Brookes.
- Solso, R. L. (2001). *Cognitive psychology*. Boston, MA: Allyn & Bacon.
- Statistics Canada. (2006). 2006 census of population. Retrieved from <http://www12.statcan.gc.ca/census-recensement/2006/index-eng.cfm>
- Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., . . . Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis*, 30, 368–388. doi:10.3102/0162373708322738
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.
- Torgerson, C. J. (2007). The quality of systematic reviews of effectiveness in literacy learning in English: A “tertiary” review. *Journal of Research in Reading*, 30, 287–315. doi:10.1111/j.1467-9817.2006.00318.x
- Torgerson, C. J., Brooks, G., & Hall, J. (2006). *A systematic review of the research literature on the use of phonics in the teaching of reading and spelling*. London, England: Department for Education and Skills. Retrieved from http://www.dfes.gov.uk/research/data/uploadfiles/RR711_.pdf
- Torgerson, C. J., & Zhu, D. (2003). A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5–16. In Research Evidence in Education Library. London, England: EPPI-Centre, Social Science Research Unit, Institute of Education.
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *The Comprehensive Test of Phonological Processing (CTOPP): Examiner's manual*. Austin, TX: Pro-Ed.
- Waks, L. J. (2007). The concept of fundamental educational change. *Educational Theory*, 57, 277–295. doi:10.1111/j.1741-5446.2007.00257.x
- Waterman, B. B. (2003). Review of the group reading assessment and diagnostic evaluation. In B. S. Plak, J. C. Impara, & R. A. Spies (Eds.), *The fifteenth mental measurements yearbook* (pp. 239–241). Lincoln, NE: Buros Institute of Mental Measurements.
- Williams, K. T. (2001). *Group reading assessment and diagnostic evaluation: Teacher's scoring & interpretive manual*. Circle Pines, MN: American Guidance Service.
- Wolgemuth, J. R., Helmer, J., Bottrell, C., Harper, H., & Lea, T. (2012). *A multi-site randomised control trial to examine the impact of ABRACA-DABRA on indigenous literacy in Australia*. Manuscript submitted for publication.
- Wolgemuth, J. R., Savage, R. S., Helmer, J., Bottrell, C., Emmett, S., Lea, T., . . . Abrami, P. (2011). Using computer-based instruction to improve indigenous early literacy in Northern Australia: A quasi-experimental study. *Australasian Journal of Educational Technology*, 27, 727–750.
- Wood, E., Specht, J., Willoughby, T., & Mueller, J. (2008). Integrating computer technology in early childhood education environments: Issues raised by early childhood educators. *Alberta Journal of Educational Research*, 54, 210–226.
- Wozney, L., Venkatesh, V., & Abrami, P. (2006). Implementing computer technologies: Teachers' perceptions and practices. *Journal of Technology and Teacher Education*, 14, 173–207. Retrieved from <http://doe.concordia.ca/cs1p/wozneyetaljtte141.pdf>

Received August 5, 2011

Revision received September 17, 2012

Accepted October 1, 2012 ■

Do Films Make You Learn? Inference Processes in Expository Film Comprehension

Maike Tibus, Anke Heier, and Stephan Schwan
Knowledge Media Research Center, Tübingen, Germany

The present article examines how suitable expository films are for learning. This question was motivated by the assumption that films are processed in a superficial manner. However, previous research has been dominated by the analyses of outcome measures and has never taken a look at online measures so that no clear conclusions have been drawn. Experiment 1 analyzed online local causal bridging inferences that are crucial for the understanding of complex scientific matters. Using a naming paradigm, it could be demonstrated that local causal bridging inferences are generated during film reception. This holds true for film viewers and audiotape listeners. Experiment 2 analyzed whether participants were able to integrate the inferred information into a coherent mental model. We found that for at least one item set, both film viewers and audiotape listeners integrated implicit information into a mental model. To further clarify the role of the pictorial information in films for the comprehension process, Experiment 3 analyzed the extent to which pictorial information can compensate for coherence breaks due to missing verbal information in the auditory channel. We found that, on a local level, pictorial information can compensate for missing verbal information, but not on a global level.

Keywords: inference processes, film comprehension, learning with films, audiovisual media

Moving pictures today dominate not only the entertainment sector in the form of big blockbusters, but they are also often used for informational purposes. Due to their realistic, vivid, experience-driven way of conveying information, videotapes and films are more frequently used as instructional media in school than other media, such as newspaper or magazine articles, computers, and video cameras (Feierabend & Klingler, 2003). A crucial precondition for conveying information successfully via films is, however, that film viewers elaborate and understand the presented content. Whereas some studies on learning outcomes of films exist with conflicting results (DeFleur, Davenport, Cronin, & DeFleur, 1992; Salomon, 1984; Weidenmann, 1989, 2002; Wetzel, Radtke, & Stern, 1994), not much is known about whether and how films are cognitively processed during the comprehension process. Analyzing local causal bridging inferences (LCB inferences) as a prototypical elaboration process, we present empirical evidence from three experiments, which demonstrate that viewers engage in elaborative processing activities during expository film comprehension.

The assumption that films are not elaborated, but rather processed superficially, has dominated common wisdom (cf. “the film watching couch potato”) and has also been supported by some research (DeFleur et al., 1992; Furnham, de Siena, & Gunter, 2002; Robinson, 1988; Salomon, 1984; Weidenmann, 1989). In particular, several studies have demonstrated that adults show

better retention and comprehension if the content is presented via text than via film (DeFleur et al., 1992; Furnham et al., 2002). Two main explanations have been proposed for this finding. First, it has been argued that due to their realism and their iconic attributes, films are perceived as “easier” to process than texts, thus evoking more shallow processing (Salomon, 1984; Weidenmann, 1989). Second, another line of argument has suggested that films present information at a high pace that cannot be controlled by the learner (Hochberg & Brooks, 1978; Robinson, 1988). It is assumed that under such conditions, the constant flow of incoming information prevents the viewer from elaborating the individual information elements appropriately. Thus, whereas readers can adapt reading speed to their cognitive processing needs, film viewers are forced to adapt their cognitive processes to the speed of presentation in the film, in turn leading to a reduction of inferences and elaboration. This echoes the argumentation of educational psychologists in the field of instructional dynamic visualizations, who have also argued that presenting audiovisual content dynamically may result in a cognitive overload of the learner and, at the same time, in suboptimal elaboration processes (e.g., Lowe, 2003, 2004; Tversky, Bauer-Morrison, & Bétrancourt, 2002).

This postulated influence of dynamic presentations on comprehension is further qualified by the fact that films present their contents simultaneously via the visual and the auditory channel. When comparing equivalent radio (i.e., audio-only information) with film (i.e., audio plus visual information) messages, results from communication studies show better comprehension of films, provided that the visual channel presents relevant information that is highly redundant with the auditory track (Drew & Grimes, 1987; Gibbons, Anderson, Smith, Field, & Fischer, 1986; Hayes, Kelly, & Mandel, 1986). Therefore, in contrast to multimedia research that has assumed either independence of the visual and the auditory channel or a negative impact of redundant messages (Mayer,

This article was published Online First December 17, 2012.

Maike Tibus, Anke Heier, and Stephan Schwan, Knowledge Media Research Center, Tübingen, Germany.

Correspondence concerning this article should be addressed to Maike Tibus, Knowledge Media Research Center, Schleichstrasse 6, 72076 Tübingen, Germany. E-mail: m.tibus@iwm-kmrc.de

2001, 2005; Sweller, 1999), several studies in the field of film comprehension have demonstrated a positive effect of audiovisual redundancy on retention and comprehension (Drew & Grimes, 1987; Grimes, 1990; Fox, 2004; Zhou, 2004). One reason for this discrepancy between multimedia research and film studies may lie in the material under investigation. Whereas multimedia research has concentrated on simplified graphical depictions that are only loosely coupled to the narration, films typically rely on dense photographic material that undergoes frequent abrupt changes via film cuts and that is strongly coupled to the narration. In the latter case, a certain amount of redundancy between visual and verbal information may render a message quicker to grasp, thereby compensating for the speed of information change. Additionally, channel redundancy may also bolster against coherence breaks within each of the information channels.

LCB Inferences as Indicators of Online Elaboration Processes

Taken together, these findings raise the question whether the dynamic character of films indeed prevents viewers from an elaborate processing of its content (as indicated by comparisons of films and texts) or whether, at least under circumstances of high audiovisual redundancy, viewers are able to engage in processes of inference and elaboration. One problem with deciding between these accounts is that, up to now, research on learning with informational films (called *expository films* in this article, i.e., films designed to convey knowledge) is characterized by a focus on outcome measures instead of process measures. That is, usually global measures of learning and comprehension are obtained subsequently to film viewing (e.g., DeFleur et al., 1992; Salomon, 1984; Wetzel et al., 1994). From these measures, propositions about cognitive processes during film viewing are derived. For several reasons, this empirical strategy is problematic. First, global outcome measures only allow the reception process of the whole film to be characterized overall, in a general way, and cannot map processing fluctuations during viewing. Second, similar learning outcomes can result from totally different viewing processes and thus represent a distorted relationship between learning process and outcome. Third, the amount of online processing may be overestimated because learners engage in additional elaboration during the subsequent testing phase, but it may also be underestimated because, due to forgetting or interference, some processing steps do not manifest during testing.

Hence, what is needed is to complement existing research on learning outcomes with the investigation of online elaboration processes in expository film comprehension. Examples for the investigation of online elaboration processes can be found both in multimedia research and in text comprehension research. In multimedia research, for example, pure outcome measures have increasingly been supplemented by process measures, such as eye-tracking data (Boucheix & Lowe, 2010; Jarodzka, Scheiter, Gerjets, & van Gog, 2010; Kriz & Hegarty, 2007). Also, the lack of online process measures in films stands in sharp contrast to the research conditions in the field of text comprehension. Here a number of valid measures of text processing have been developed, which have given way to corresponding elaborate models of text comprehension that have also been generalized to other media, including films (Graesser, Wiemer-Hastings, & Wiemer-Hastings,

2001; Magliano, Dijkstra, & Zwaan, 1996; Magliano, Miller, & Zwaan, 2001; Zwaan & Radvansky, 1998). In particular, it is argued that although films and texts are different media using different symbol systems, they still follow similar higher order principles of structuring and presenting information (Gernsbacher, Varner, & Faust, 1990; Graesser et al., 2001; Magliano et al., 2001; Zwaan & Radvansky, 1998). Accordingly, it has been shown that structural equivalent texts and films are recalled in a similar manner (Baggett, 1979).

Basically, the models assume that readers try to understand the text by generating a coherent mental representation of its content. This is done not only by appropriately encoding the textual information but also by generating additional inferences, which both link the different parts of the content and relate them to already existing knowledge structures (e.g., Fincher-Kiefer & D'Agostino, 2004; Graesser, Millis, & Zwaan, 1997; Graesser, Singer, & Trabasso, 1994; Kintsch, 1988, 1998; Long, Golding, & Graesser, 1992; Wiley & Myers, 2003). In other words, inferences, which can be broadly defined as information that is not explicitly stated but is generated by the recipient himself from the given information,¹ are crucial for text understanding. One type of inference, which is of particular importance for understanding complex scientific matters, is the LCB inference. LCB inferences establish causal coherence on a semantic or conceptual level in order to generate a coherent mental model (Graesser, León, & Otero, 2002). As they are essential for comprehension, they are generated online, that is, during the reception process (e.g., Graesser et al., 2002, 2007). However, causal bridging inferences are not reliably generated in expository text comprehension (e.g., Graesser & Bertus, 1998; Millis & Graesser, 1994; Noordman, Vonk, & Kempff, 1992; Singer, Harkness, & Stewart, 1997; Wiley & Myers, 2003). Instead their generation depends to a large degree on, for instance, an instruction or a reading goal that encourages comprehension processes, the availability of the necessary information to compute the inferences, and sufficient prior knowledge of the recipient to relate to the new information. However, to elaborate the presented content adequately, the reader nevertheless needs to generate LCB inferences. Thus, the generation of LCB inferences can be regarded as a valid indicator for relevant elaboration activities during expository text comprehension. Consequently, by combining the prominent role of LCB inferences for comprehension of expository texts with the notion that text comprehension models are generalizable to films, as in the present article, it can be argued that LCB inferences can serve as indicators for online elaboration activities in film comprehension.

Experimental Overview

In order to determine whether viewers elaborate on expository films, three experiments were conducted. In Experiment 1, by applying a naming task paradigm (Long et al., 1992; Murray & Burke, 2003; Potts, Keenan, & Golding, 1988), we investigated to what extent LCB inferences are generated online during expository film comprehension. In Experiment 2, we examined the role of online elaboration for outcome comprehension by administering a comprehension test referring to the LCB inferences that, according

¹ For better readability, the masculine form is used in this article but refers to both masculine and feminine.

to Experiment 1, were inferred during film viewing. In both experiments, presentation of the film was compared with an audio-only presentation in order to differentiate between the impact of presenting information dynamically and the effect of giving information simultaneously via both auditory and visual channels. The role of the visual channel for elaboration was further investigated in Experiment 3. Here we analyzed to what extent pictorial information can compensate for coherence breaks due to missing verbal information in the auditory channel.

Experiment 1

In Experiment 1, we analyzed whether LCB inferences as indicator for online cognitive processing are generated in expository film comprehension via a naming paradigm. In addition, the effect of the dynamic pictorial information on the generation of LCB inferences was tested.

We hypothesized that if participants generate LCB inferences during expository film comprehension, then naming latencies for words that represent the respective inferences should be shorter in an inference-related (IR) context than in an inference-unrelated (IU) context. When presented in the IR context, the naming word represents the required LCB inference. If participants generate the required LCB inference during the comprehension process, then the generated LCB inference activates the respective concept in the semantic network. This activation serves as a prime and facilitates naming latencies for a word that represents the generated inference. In contrast, when presented at another position in a second film where the same word does not represent a generated inference (i.e., in an IU context, which serves as a control measure), naming latencies should not be facilitated.

Two central features of films may contribute to the occurrence or nonoccurrence of LCB inferences. Namely, films present their content dynamically, and, in films, information is distributed simultaneously via two information channels. If dynamic presentation per se hinders viewers from generating inferences due to constraints in available processing time, no indications of LCB inferences should show up, irrespective of number of information channels. In contrast, if dynamic presentation is in principle no obstacle to comprehension, evidence of LCB inferences should be observed. Furthermore, if the impact of dynamic presentation on inference generation is qualified by the existence of the pictorial information via the visual channel, this should be evident in comparison with an audio-only condition. In particular, if the addition of redundant pictorial information further increases the demands for information processing (Mayer, 2001, 2005; Sweller, 1999), indications of LCB inferences should be observed in the audio condition, but not in the video condition. Instead, if redundant information from the visual channel contributes positively to understanding (Drew & Grimes, 1987; Gibbons et al., 1986; Hayes et al., 1986), indications of LCB inferences should be observed in the video condition, but not in the audio condition.

Method

Participants. Sixty-three undergraduate students (47 women, 16 men) from the University of Tübingen, Germany, participated in this experiment. Their average age was 24.32 years ($SD = 3.56$). Participants received either course credit or payment for participation.

Materials. The film material was taken from "Sendung mit der Maus" ("The Mouse Show"; Maiwald & Biermann, 2006a, 2006b, 2006c), a popular, award-winning German educational TV program, for two reasons: First, it is a TV program specifically designed for a younger audience and therefore is characterized by a high degree of audiovisual redundancy (Walma van der Molen & van der Voort, 2000). Second, although the target audience for "Sendung mit der Maus" are primary school students, the average age of its viewers is 39 years. From a set of film clips from the TV show, three clips were chosen that demonstrated several causal steps to make LCB inferences necessary showing substantial overlap between pictorial and verbal information and that were long enough to provide a complex content. One film clip ("Rear View Mirror"; 4.17 min.; Maiwald & Biermann, 2006a) served as a trial movie to acquaint participants with the task and situation. One film clip about the "construction of a thermos" (7.19 min.; Maiwald & Biermann, 2006c) and one about the "formation of lightning" (6.05 min.; Maiwald & Biermann, 2006b) served as experimental films.

In the expository film, "the construction of a thermos," the protagonist tries to keep his tea warm over time. He starts by discovering that different materials have different heat conduction. He then detects that air has an insulating effect on heat and that it is helpful to put a lid on top of a glass container to keep the tea inside the glass warm. Next, he infers that aluminum foil reflects the heat and is, therefore, a good heat containing enclosure. At the end of the film, he discovers that a real thermos is built according to the heat preserving ideas that he came up with. The expository film "formation of lightning" starts by demonstrating the effect of generating electricity by rubbing a sweater on a balloon. Then the effect that a balloon loaded with electricity attracts hair is demonstrated. Subsequently, it is shown that a balloon has a limited storage capacity. These principles are then transferred to the formation of lightning and exemplified by photographing real lightning with a special camera.

Preparation of the experimental material. As a first step, a propositional analysis (Bovair & Kieras, 1985; Kintsch, 1998) was applied to transcripts of audio traces of the two experimental film clips (thermos, lightning). This was done in order to identify possible positions in the transcript that could be manipulated in a way that made LCB inference generation necessary for the recipient to understand the content. These positions are called *LCB inference positions* in the following. After transcribing the original audio traces for each experimental film, propositional analyses were conducted for the film transcripts by two trained analyzers individually. The analyzers agreed with 88% of the propositions for the thermos film and with 90% of the propositions for the lightning film. Conflicts were resolved by discussion.

As an example of an LCB inference position, in the "lightning" film, the protagonist demonstrates the concept of unloading electricity by means of an example. The narrator says:

... And when there is lightning? For that, we take our balloon again. But this time, we don't rub electricity onto it. We fill it with water. That's fine for a while, but, at some point, the balloon's capacity is reached and the balloon bursts. You can also say that the balloon bursts "lightning-fast." And if the sky can't store any more electricity, it "unloads" electricity in the form of lightning."

Here, the conclusion of the last sentence can be regarded as an LCB inference that relates the observed phenomenon (lightning) to

its underlying cause (unloading of electricity). Overall, in the "thermos" film, eight LCB inference positions and in the "lightning" film, six LCB inference positions (i.e., 14 naming positions across both films) were identified, relating to conductance, insulation, voltage, or friction, for example.

Next, again based on the results of the propositional analysis, naming words were identified. Typically, they consisted of a noun or a verb that was the central element of the respective sentence in the narration and were supposed to best reflect the respective LCB inference that needed to be generated in order to make the film coherent. Thus, in the example above, the German verb *entladen* (to unload) was selected as a naming word. To ensure the validity of the naming words, two prior studies were conducted that showed that participants had both enough prior knowledge and tended to generate spontaneously the respective naming words as inferences at the respective LCB inference positions in the film.

As a third step, both for the thermos and the lightning film, two film versions were constructed according to the following procedure: The possible LCB inference positions were split in half. For one half of the possible LCB inference positions, the causal information that bridged one piece of information to another piece of information was omitted from the film transcript so that the respective causal information needed to be generated by the recipient in the form of an LCB inference. The other half of the possible LCB inference positions was left unchanged. This procedure resulted in two complementary versions of the thermos film, with each including four LCB inference positions at which an inference was necessary; the remaining four positions served as an IU context for the naming words of the "lightning" film, and therefore no inference was necessary. A complementary procedure was created for the "lightning film." For the experiment, the films were grouped in two pairs, namely, Version 1 of the thermos film together with Version 1 of the lightning film, and Version 2 of the thermos film together with Version 2 of the lightning film. Each participant saw one of these film pairs, with the presentation order being balanced across participants.

Therefore, in the above-mentioned "lightning film" example, the LCB inference information of the mentioned paragraph ("it unloads electricity") was omitted and instead *entladen* (to unload) was inserted as naming word. By means of this procedure, the IR context for *entladen* was established in Version 1 of the "lightning film." Additionally, the same word was also inserted at an unedited LCB inference position in Version 1 of the "thermos film," thereby serving as an IU context. Hence, every naming word appeared twice during the experiment, namely, either in an IR context at the respective LCB inference position in the "lightning film" and, correspondingly, in an IU context at one of the inference positions in the "thermos film," or vice versa. The two films served therefore with as mutual control contexts.

Finally, by deletion of the visual trace of the films, a corresponding set of audio presentations was generated.

Other materials distributed. During the experiment, participants' interest in physics and chemistry (five 4-point Likert scale questions), topic-related knowledge (six multiple-choice questions), perceived cognitive load (NASA-TLX [Task Load Index]; Hart & Staveland, 1988), and working memory span (German version of the "reading span task" of Daneman & Carpenter, 1980) were measured.

To ensure that participants tried to understand the content of the films (important for an LCB inference generation in expository domains), participants answered two *comprehension test questions* regarding the film content after watching the film. Participants were instructed to try to understand the presented content and were informed that a comprehension test would be administered after the respective film. The comprehension scores were not analyzed further. This is in line with other experiments in inference research (e.g., Albrecht & O'Brien, 1993).

To prevent carryover effects from the first to the second film, participants rated 10 more or less well-known art paintings regarding their appearance, the artistic composition, and so forth, on seven dimensions. The *picture-rating test* was not analyzed further.

To evaluate the obtrusiveness of the naming task, participants answered two questions regarding *how disruptive* they experienced the *naming task* to be. Items were answered on a 4-point Likert scale ranging from 1 (*very low*) to 4 (*very high*). Overall naming task disruption was averaged across these two questions.

Variables and design. Two independent variables were included in this experiment in a 2×2 mixed factorial design: *context* and *presentation mode*. The context (IR vs. IU), in which the naming task was presented, was varied as a within-subjects factor to obtain naming latencies for the same word twice by the same person: once when the respective LCB inference was required (IR context) and once when the respective LCB inference was not required (IU context). Furthermore, two presentation modes (video vs. audio) were administered in this experiment as a between-subjects factor to analyze the effect of the pictorial information on LCB inference processes. The audio condition comprised the same verbal information as the video condition, but lacked the pictorial information. Additionally, different naming task word sets were varied across participants, but solely for methodological reasons. Therefore, word set is not considered as an experimental between-subjects factor. However, to control its effect statistically, it was included in the respective analyses.

The dependent variable was the *naming latencies* for the naming words. Because both contexts were presented as a balanced within-subjects factor, it was ensured that interindividual and word-related differences (such as word frequency, number of syllables, etc.) could not bias the facilitation effect caused by an LCB inference generation.

Procedure. The experiment was run computer-controlled and individually. The experimenter was present during the experiment. He welcomed the participant and led him to a desk with a computer and a monitor. Participants started the experiment by reading the instructions and answering questions regarding their interest in physics and chemistry. Following this, participants stated their demographic data. Then the instructions for the trial film and the trial film were presented. Participants listened to all films or audiotapes via headphones. For better readability, the term *film* is used for the film and the audiotape in the following context because the procedure was the same for both conditions. Subsequently, the instructions for the experimental film were presented, followed by the first experimental film. The experimental procedures were controlled by a Microsoft computer and programmed using MediaLab and directRT. Film clips were presented at full screen on a 17-in. monitor.

For each word naming task, the film was interrupted by showing a blank screen for 250 ms, followed by an asterisk for 500 ms. Then the naming word (e.g., *entladen/to unload*) was presented. The participants' task was to name the word aloud as quickly as possible. A voice-activated relay was used to determine the latencies between the presentation of the word and the participants' response. In addition, an experimenter monitored the session and recorded any pronunciation errors. When participants named the word aloud, the word disappeared and the film restarted where it had left off. Audiotape listeners looked at a black monitor screen while listening to the audiotape. The experimenter ensured that participants focused on the screen throughout the experiment as the naming tasks appeared randomly and were not announced aurally.

Subsequent to the film, all participants answered the cognitive load items, followed by the comprehension test questions. Next, participants took the picture-rating test that lasted approximately 10 min. Then participants proceeded with the second film. The procedure was identical to the first film. After answering the comprehension test questions regarding the second film, participants rated how disruptive the word-naming task was overall. Subsequently, participants completed a topic-related knowledge test and, finally, a working memory span test.

Results and Discussion

Control variables. To ensure that the control variables (interest, topic-related knowledge, age, working memory span) were equally distributed across the conditions, we conducted four 2 (presentation mode) \times 2 (word set) analyses of variance (ANOVAs). The analyses revealed that all control variables were equally distributed across conditions [all F s < 1 , except for working memory span across presentation modes), $F(1, 59) = 2.59$, $MSE = .45$, $p > .10$, $\eta_p^2 = .04$; age across word sets, $F(1, 59) = 1.19$, $MSE = 12.76$, $p > .10$, $\eta_p^2 = .02$; and *working memory span* across word sets, $F(1, 59) = 14.99$, $MSE = .45$, $p < .001$, $\eta_p^2 = .20$. Participants who received Word Set 2 in the naming task had a higher working memory span than participants who received Word Set 1 in the naming task. However, no correlations between working memory span and naming latencies' differences were found ($r = -.01$). Hence, we conducted no further analyses.

Naming latencies. There were virtually no pronunciation errors (one word out of 882 = 0.11%). Naming latencies greater than 2000 ms were treated as missing data, which constituted 0.68% of the data. The naming latencies followed a Gaussian distribution. We performed a 2 (context) \times 2 (presentation mode) \times 2 (word set) ANOVA on the naming latencies. The analysis revealed a highly significant main effect of context on the naming latencies, $F(1, 59) = 41.19$, $MSE = 1727.42$, $p < .001$, $\eta_p^2 = .41$. As expected, naming latencies in the IR context were shorter than naming latencies in the IU context (530 ms vs. 578 ms; see Table 1). This was the same across presentation modes and word sets (all F s < 1). Hence, the results demonstrate that participants both in the video and in the audio condition showed the expected facilitation effect of context on their naming latencies to a similar extent, indicating generation of LCB inferences in expository film and audiotape comprehension. This finding suggests that the presentation of dynamic information per se does not necessarily hinder processes of inference and elaboration during reception.

Table 1

Means (and Standard Deviations) of Naming Latencies (in Milliseconds) as a Function of Context, Presentation Mode, and Word Set

Context	Presentation mode	Word set	<i>M</i> (<i>SD</i>)	<i>N</i>
IR context	Video	1	505 (89)	18
		2	522 (82)	15
		Overall	513 (85)	33
	Audio	1	544 (119)	15
		2	555 (119)	15
		Overall	549 (117)	30
	Overall	1	523 (104)	33
		2	538 (102)	30
		Overall	530 (103)	63
IU context	Video	1	561 (117)	18
		2	561 (93)	15
		Overall	561 (105)	33
	Audio	1	607 (143)	15
		2	588 (118)	15
		Overall	597 (129)	30
	Overall	1	582 (129)	33
		2	574 (105)	30
		Overall	578 (118)	63

Note. IR = inference-related; IU = inference-unrelated.

Furthermore, the presence or absence of a visual channel had no substantial influence on this observation. Hence, there was no indication of a redundancy effect, neither in its positive formulation that redundant information fosters comprehension by dual coding (Fox, 2004) nor in its negative formulation that redundant information hinders comprehension by an unnecessary increase in mental load (Sweller, 1999).

Cognitive load and naming task disruption. To analyze the effects of presentation mode and word set on measures of participants' experienced cognitive load and participants' experienced overall disruption of the naming task, we conducted two 2 (presentation mode) \times 2 (word set) ANOVAs. Due to technical problems, the measures of the first 14 participants were not recorded for these two variables. This affected all conditions comparably. The total sample for these analyses was reduced to 49. Analyses did not reveal any effects for presentation mode, nor for word set on all three cognitive load measures (all F s < 1). Analyses did not reveal any effects for presentation mode, nor for word set on the overall disruption of the naming task (all F s < 1). Hence, in line with the findings on naming latencies, no indication was found that the pictorial information in films demands more cognitive resources than a solely verbal presentation. Overall, participants across presentation modes rated the naming task as "not very disrupting," confirming the unobtrusiveness of the naming task.

Experiment 2

Experiment 1 provided evidence that LCB inferences are generated online in expository film comprehension. On the basis of these findings, the question arises how LCB inferences generated online relate to learning outcomes. An answer to this question can be found on the basis of a comprehension test that specifically addresses those positions in the films that have been identified as

LCB inference positions in the previous experiment. By comparing a condition in which LCB inferences have to be generated at the respective positions in the film (implicit information condition) with both a condition in which the relevant information is explicitly given (explicit information condition) and a control condition in which no information prior to the comprehension test is given, the assumption that LCB inferences are permanently integrated into a mental model of the film content was tested. We hypothesized that if, on the one hand, the online-generated LCB inferences are a transient processing step during online reception, but are not integrated into a coherent mental model that is permanently stored in long-term memory, then performance on a comprehension test that is given subsequently to film viewing should be substantially lower in the implicit condition compared with the explicit condition. In which case, it would more or less equate with a "no-information" or control condition. In contrast, if the online-generated LCB inferences normally lead to permanent integration of inferred information into the mental model of the film content, then participants in the implicit information condition should substantially outperform the control condition. Their performance should be comparable to the condition in which the relevant information is explicitly stated.

In addition, the interplay of dynamic presentation and number of information channels was addressed similarly to Experiment 1. More specifically, we hypothesized that if the addition of redundant pictorial information to the mere audio presentation further increases the demands for information processing, indications of a permanent integration of LCB inferences should be observed in the audio condition, but not in the video condition. Instead, if redundant pictorial information from the visual channel contributes positively to understanding, indications of a permanent integration of LCB inferences should be observed in the video condition, but not in the audio condition.

Method

Participants. Ninety-four undergraduate students (67 women, 27 men) from the University of Tübingen, Germany, participated in this study. Their average age was 23.43 years ($SD = 3.46$). The participants received either payment or course credit for participation in the experiment.

Materials. The basic film material (film, audiotape) was identical to the experimental material used in Experiment 1. However, in the present experiment, the film² was presented continuously and was not interrupted by naming tasks. Nevertheless, as in Experiment 1, the eight positions of the thermos film and the six positions of the lightning film that had been identified as possible positions for LCB inference were treated similarly to Experiment 1. That is, for one half of the possible LCB inference positions, the causal information that bridged one piece of information to another piece of information was omitted from the film transcript so that the respective causal information needed to be generated by the recipient in the form of an LCB inference. The other half of the possible LCB inference positions was left unchanged. That is, it included the causal information so that it was not necessary for the recipient to generate LCB inferences. Two film versions of both the thermos and the lightning film were created so that sequences starting with implicit information or starting with explicit information were counterbalanced.

Additionally, a *comprehension test* was constructed and discussed with three other people with expert knowledge in the domain. It consisted of 13 items, whereby seven items addressed the "construction of a thermos," and six items addressed the "development of lightning." Each item covered a position that had been identified as an LCB inference position in Experiment 1. The items were presented in a multiple-choice format: one correct answer plus three distractors. The correct answer reflected the LCB inference generated online (measured in Experiment 1) plus its integration with long-term memory structures into a coherent mental model. Moreover, the correct answer contained the naming word from Experiment 1 (the naming word reflected the LCB inference that had to be drawn at the respective position). To avoid word-based memory effects, one of the three distractors also contained the naming word from Experiment 1. The items' order of appearance in the comprehension test corresponded to their order of appearance in the films. In the case of the example already described for Experiment 1 (unloading of electricity), the respective comprehension test item reads as follows:

When rubbing on the air (molecules), falling raindrops get charged with electricity. Thereby, lightning may be generated. Which one of the four following explanations for the generation of lightning is correct? (1) (correct answer) The raindrop's storage capacity for electricity is limited and when the limit is reached, they unload in form of lightning. (2) Falling raindrops increase their volume. When they have reached a critical size, they divide and lightning is generated. (3) As soon as falling raindrops go below a critical distance from earth, they unload in the form of lightning. (4) Falling raindrops heat up when they reach warmer air layers. When a critical temperature and therewith a critical intensity of molecule movement is reached, lightning is generated.

To measure participants' interest in physics and chemistry, cognitive load, and topic-related knowledge, participants answered the same items as in Experiment 1.

Variables and design. Presentation mode was varied as a between-subjects factor, and participants were randomly assigned to the video, audio, or control conditions. Participants in the video condition viewed the film normally, whereas participants in the audio condition listened to the audio trace of the film but watched a black screen instead of the visual trace. Participants in the control condition answered the comprehension questions without receiving any experimental material (i.e., film, audiotape). Additionally, within both the video and audio conditions, two presentation variations were implemented, which counterbalanced the LCB inference positions that were presented explicitly or implicitly. This led to a between-subjects factor (presentation mode) with five levels: video_{Variant1}, video_{Variant2}, audio_{Variant1}, audio_{Variant2}, control. Additionally, as a two-level within-subjects factor, the subsequent comprehension test was split in two sets of items. Item Set 1 consisted of questions addressing LCB inferences that provided explicit information in video_{Variant1} and audio_{Variant1}, whereas they provided implicit information in video_{Variant2} and audio_{Variant2}. No prior information related to these items was given in the control condition. Complementarily, Item Set 2 consisted of questions addressing LCB inferences that provided explicit information in

² As in the other experiments, the term *film* is used for the film and the audio trace for a better readability.

video_{variant2} and audio_{variant2}, whereas they provided implicit information in video_{variant1} and audio_{variant1}. Again, no prior information related to these items was given in the control condition. The main dependent variable was the score of correctly answered items in the comprehension test.

Procedure. The experiment was run quasi-individually. Participants in groups of six were tested together in one room with six desks that were separated by partitions. Each participant worked at his or her own HP Compaq laptop with a 17-in. monitor. The experimenter was present during the whole experiment. Once all participants had arrived, the experimenter started the experiment individually at each participant's laptop, one after another. Each participant listened to the experimental material with individual headphones. The experiment started by asking five items about the participant's interests in physics and chemistry. Afterward, participants provided their demographical data. Then the first film started. Participants were instructed to follow the film attentively and to try to comprehend the content. It was announced that, after the film comprehension test, questions regarding the film content would follow. After the film, participants first answered the comprehension questions and, second, the cognitive load items. Then the second film started and participants proceeded in the same manner. After answering the comprehension and cognitive load items regarding the second film, participants took the topic-related knowledge test. When everyone was finished, participants received their money or course credit and left. As in Experiment 1, the films' presentation order and the film versions were counterbalanced.

The material and procedure for the control condition were identical to the experimental conditions, except that participants in the control condition did not receive any experimental material and thus also did not answer cognitive load items. The presentation order of the comprehension test items was counterbalanced. Half the participants in the control condition answered comprehension test items regarding the construction of a thermos first; the other half answered comprehension test items regarding the formation of lightning first.

Results and Discussion

Control variables. To ensure that the control variables (interest, topic-related knowledge, age) were equally distributed across conditions, we performed several ANOVAs. In a first step, we performed three one-factorial ANOVAs. The analyses revealed that *interest* and *age* were equally distributed across the presentation modes (all $F_s < 1$). Topic-related knowledge, however, was not distributed equally across presentation modes, $F(2, 91) = 3.65$, $MSE = 342.70$, $p < .05$, $\eta_p^2 = .07$. Post hoc analyses with a Bonferroni test revealed that participants in the audio condition had significantly less topic-related knowledge than participants in the video condition and than participants in the control condition ($p < .05$).

Comprehension scores. To investigate whether participants integrated the inferred information into long-term memory structures, we conducted a two-factorial analysis of covariance, with comprehension performance as the dependent variable. The between-subjects factor presentation mode included five conditions, namely video_{variant1}, video_{variant2}, audio_{variant1}, audio_{variant2}, and the control condition. The within-subjects factor item set

defined which subset of questions in the comprehension test addressed explicitly or implicitly given information at LCB inference positions. We included topic-related knowledge as a covariate because it was not equally distributed across presentation modes.

The analysis yielded a significant main effect for presentation mode, $F(4, 88) = 10.47$, $MSE = 164.51$, $p < .001$, $\eta_p^2 = .32$, and a significant interaction between presentation mode and item set, $F(4, 88) = 3.43$, $MSE = 218.34$, $p < .05$, $\eta_p^2 = .14$. Also, topic-related knowledge was significant as covariate, $F(1, 88) = 7.49$, $MSE = 164.48$, $p < .01$, $\eta_p^2 = .08$. Post hoc Bonferroni analyses revealed the following pattern (see Table 2): Both for the video as well as the audio condition, no differences were found between the comprehension scores for LCB inference positions with implicitly given information versus LCB inference positions with explicitly given information.

Regarding the differences between implicit, explicit, and no information, we found the following pattern: For Item Set 2, all experimental conditions led to significantly higher comprehension scores than the control group. This result is in line with the hypothesis that the recipients of the implicit condition inferred the implied information and integrated it into long-term memory structures so that it was accessible during the subsequent comprehension test. For Item Set 1, however, contrary to our expectations, only the participants in the conditions in which the relevant information was explicitly given outperformed participants in the control group, whereas the participants in the conditions in which the information was implicitly given information did not differ significantly from the control group. This pattern was true for participants in the video as well as the audio conditions. It seems as if for this item set, participants who had to infer implicit information were not able to integrate the inferred information into long-term memory structures, as were those participants who received explicit information. The fact that this pattern was only found for this item set and not for both given item sets leads us to the suggestion that this finding may be specific to Item Set 1. A closer look at Item Set 1 reveals comparatively high comprehension scores indicating that this item set was comparatively easy to solve. It may be that, because this material was relatively easy, participants in the implicit condition did not put enough effort into the inference-making process. Their scores, therefore, did not differ from the control group's scores.

Finally, similar to Experiment 1, we found no differences between the video condition and the audio condition. In other words, the pictorial information via the visual channel did not influence the information-processing process or the resulting construction of the mental model.

Cognitive load. To analyze the effects of presentation mode on participants' experienced *cognitive load measures* (effort, con-

Table 2
Means (and Standard Deviations) of Comprehension Scores as a Function of Presentation Mode and Item Set

Presentation mode	Item Set 1	Item Set 2
Video (explicit)	81.25 (13.44)	76.78 (13.68)
Video (implicit)	78.13 (15.77)	66.96 (14.49)
Audio (explicit)	82.29 (16.63)	68.75 (21.65)
Audio (implicit)	70.83 (19.72)	67.86 (17.69)
Control	66.11 (18.81)	47.62 (16.06)

fidence, stress), we conducted three one-factorial ANOVAs. Analyses did not reveal any effects (all $F_s < 1$).

Experiment 3

Experiments 1 and 2 provided evidence that viewers generate LCB inferences during film viewing and that these inferences are integrated into their respective mental models. Thus, the dynamic and simultaneous information presentation was no principled obstacle for an elaborate processing of the to-be-learned content. However, generation of LCB inferences was evident both for film and audio-only presentation. This raises the question as to what extent film pictures contribute to the inference-generating process or whether they just have a decorative function. One reason for the lack of differences between the audiovisual and the audio-only presentation in the previous two experiments could be that, according to the propositional analysis of the material, the narration was highly coherent. In this case, the audio track may be sufficient for comprehension. However, although both high coherence and high redundancy are typical of educational films aiming at children as the target audience, other educational films or documentaries may not be as coherent. In these cases, the visual channel may compensate for incoherence in the narration. We tested this hypothesis in Experiment 3 by deliberately including coherence breaks in one of the films used in the previous two experiments.

Method

Participants. Sixty students (39 women and 21 men) from the University of Tübingen (Germany) took part in the experiment. Their average age was 25.47 years ($SD = 3.41$). The participants received either payment or course credit for participation in the experiment.

Materials. Two versions of the “thermos” film (Maiwald & Biermann, 2006c) were used in the experiment, which varied with respect to the coherence of the film’s narration. The narration of the highly coherent film version was largely identical with the original film version, except for five reformulations where incoherencies were identified by the propositional analysis described in Experiment 1. Here, the original text was replaced by a more explicit and therefore more coherent formulation. For example, at one point in the film, the protagonist puts several different materials, namely, pieces of wood, glass, rubber, and metal into a pot of hot water in order to demonstrate their difference in heat absorption. Although the materials are clearly visible in the film, they are not mentioned explicitly in the narration. Therefore, for the coherent version, a sentence was included that names the respective materials explicitly and explains their differences in heat absorption.

In contrast, for the incoherent version, at 15 positions in the film’s narration, a coherence break was inserted into the audio track that could be compensated for by attending to the accompanying pictorial information. For example, at one point in the film, the protagonist demonstrates the principles of heat reflection by putting butter into a normal glass and into a glass wrapped with aluminum foil, and irradiates both glasses with an infrared lamp. The accompanying narration says: “Now the butter test. Christoph puts a piece of butter into each glass. A clear result: In the glass without aluminum foil, the butter melts. In the glass wrapped with

aluminum foil, the butter does not melt because the heat is reflected by the aluminum foil” (Maiwald & Biermann, 2006c). In order to create an incoherent narration, the part “because the heat is reflected by the aluminum foil” was taken out. However, by attending to the accompanying moving picture that shows that aluminum foil has a silvery surface that can reflect the heat, this incoherence can be compensated for.

Next, a comprehension test was developed. The comprehension test consisted of three parts: (a) a *summary task* (“Please describe the setup of a thermos and how it functions”); (b) a *verification task* that consisted of 12 statements. Six statements related to positions that were incoherent in the narration of the incoherent version, but coherent in the narration of the coherent version. Half of these statements were true and the other half were false. For example, regarding the “heat absorption” example described above, the corresponding (false) item of the comprehension test was: “Rubber absorbs more heat than glass.” The other six statements related to positions that were coherent in the narration of all versions, with half of the statements being true and the other half being false; (c) a *transfer test* that consisted of three questions: “How should a cool bag be designed?” (near transfer), “How should one be dressed to keep warm as long as possible at cold temperatures?” (far transfer), “Which measures should be taken to heat a liquid as quickly as possible?” (far transfer).

Additionally, as in Experiments 1 and 2, participants’ interest in physics and chemistry, topic-related knowledge, and cognitive load were measured.

Procedure and design. Two independent variables, namely, coherence of narration (coherent vs. incoherent) and presentation mode (video vs. audio) were included in this experiment in a 2×2 between-subjects design. The experiment was run in computer-controlled form in individual sessions. The procedure was identical to Experiment 2, except that only the thermos film was presented. After the film or the audio presentation, participants answered the comprehension questions.

Data analysis. The summaries were analyzed on the basis of an expert solution in which a total of 10 relevant facts and causal connections were specified. Accordingly, for each summary, the number of facts and causal connections that were mentioned was determined. Similar rating schemes were developed for the three transfer questions. The interrater agreement was determined according to Cohen’s kappa, with $\kappa = .91$ for the summaries, and $\kappa = .96$, $\kappa = .99$, and $\kappa = .85$ for the three transfer questions, respectively.

Results and Discussion

Control variables. According to separate 2 (presentation mode) $\times 2$ (coherence of narration) ANOVAs, there were no a priori differences between the experimental groups with regard to age, interest in physics and chemistry (all $F_s < 1$), and topic-related knowledge, $F(3, 56) = 1.92$, $p > .1$.

Comprehension scores. We analyzed the results of the comprehension tests in two separate 2 (coherence of narration) $\times 2$ (presentation mode) ANOVAs. The sum score of the six items related to those positions that were incoherent in the narration of the incoherent versions, but coherent in the coherent versions, showed a significant effect of coherence of narration, $F(1, 56) = 5.40$, $MSE = 1.00$, $p < .05$, $\eta_p^2 = .09$. Significant effects also

emerged for presentation mode, $F(1, 56) = 8.07$, $MSE = 1.00$, $p < .01$, $\eta_p^2 = .13$, and the interaction of both factors, $F(1, 56) = 4.27$, $MSE = 1.00$, $p < .05$, $\eta_p^2 = .07$. Post hoc analyses using a Bonferroni test revealed that participants who heard the incoherent version performed significantly worse than participants in all other conditions. In contrast, the other conditions did not differ from each other (see Table 3). Thus, in correspondence with the results of the first two experiments, participants in the video and audio versions were able to answer respective comprehension items equally well when key parts of the narration were presented in coherent ways. However, more importantly, a substantial difference between audio presentation and video presentation was found if a certain part of the narration was presented in an incoherent manner. For example, although the comprehension performance was substantially reduced for the audio presentation of the incoherent narration, this drop in performance was not found in the video condition. Here, the viewers were able to compensate for a coherence break in the narration by attending to the accompanying moving picture. In other words, the results indicate that the pictorial information in videos does not merely serve an illustrative function, but instead plays an important role in compensating for coherence breaks in verbal information. These “visual inferences” are crucial for creating a coherent mental representation.

In contrast, with regard to the sum score of those six items related to positions that were coherent in all versions, we found no significant effects or interactions (all $F_s < 1$). This finding indicates that the observed effects of incoherence on comprehension described above were not due to a reduced coherence of the incoherent version in general. Instead, these effects were specific to local positions in the narration where a coherence break occurred.

We analyzed the summary scores in a 2 (coherence of narration) \times 2 (presentation mode) ANOVA. The results show a significant effect of the coherence of narration, $F(1, 56) = 14.3$, $MSE = 7.80$, $p < .01$, $\eta_p^2 = .20$. Both in the film and in the audio condition, the versions with a coherent narration led to a more complete summary than the versions with an incoherent narration ($M = 5.53$, $SD = 1.19$ vs. $M = 3.67$, $SD = 2.19$ for the video conditions; $M = 5.40$, $SD = 1.60$ vs. $M = 4.00$, $SD = 1.56$ for the audio conditions). Neither the factor presentation mode nor its interaction with coherence of narration was significant (with both $F_s < 1$). Thus, although the participants were able to compensate for coherence breaks in the narration by attending to the accompanying moving pictures on a local level, their overall comprehension of the film was nevertheless disturbed.

We analyzed the three transfer test scores in separate 2 (coherence of narration) \times 2 (presentation mode) ANOVAs. The first transfer task asked the participants to describe the construction of

a cool bag. Here, a main effect of coherence of narration was found, $F(3, 56) = 4.68$, $MSE = 1.42$, $p < .05$, $\eta_p^2 = .08$. Again, both in the film and in the audio condition, the version with a coherent narration led to a more complete solution of the transfer task than the version with an incoherent narration ($M = 2.93$, $SD = 1.39$ vs. $M = 2.13$, $SD = 1.13$ for the video conditions; $M = 2.67$, $SD = 1.12$ vs. $M = 2.13$, $SD = 1.13$ for the audio conditions). Neither the factor presentation mode nor its interaction with coherence of narration was significant (with both $F_s < 1$). For the other transfer tasks, no significant effects were found (all $F_s < 1$).

Cognitive load. To analyze the effects of presentation mode and coherence of narration on participants’ experienced cognitive load measures (effort, confidence, stress), we conducted three 2 \times 2 ANOVAs. Analyses revealed a significant effect for coherence of narration, $F(3, 56) = 7.50$, $MSE = 0.14$, $p < .01$, $\eta_p^2 = .12$, and for presentation mode, $F(3, 56) = 5.74$, $MSE = 0.14$, $p < .05$, $\eta_p^2 = .09$, on the Effort scale. For the Confidence scale, we found a significant effect for the factor presentation mode, $F(3, 56) = 9.74$, $MSE = 0.19$, $p < .01$, $\eta_p^2 = .15$ (all other $F_s < 1$). For the Stress scale, we found no significant effects (all $F_s < 1$ except for presentation mode), $F(3, 56) = 3.71$, $MSE = 0.22$, $p > .10$, $\eta_p^2 = .06$. In other words, participants in the video condition invested more effort ($M = 2.62$, $SD = .39$ for video vs. $M = 2.38$, $SD = .41$ for audio) as well as participants in the incoherent condition ($M = 2.63$, $SD = .43$ for incoherent vs. $M = 2.37$, $SD = .35$ for coherent). Participants in the audio condition, however, experienced more confidence ($M = 1.88$, $SD = .41$ for audio vs. $M = 1.53$, $SD = .45$ for video).

Taken together, the findings of Experiment 3 show that the pictorial information is not merely an illustration of the narration, but instead can play an important role in comprehension. In particular, the existence of pictorial information seems to help viewers to get across coherence breaks in the auditory narration. However, as the results of the summaries and the near transfer task show, this fostering function is restricted locally. Overall, a narration that includes a substantial number of coherence breaks (15 in the present case) decreases comprehension, irrespective of the existence of pictorial information presented via the visual channel.

General Discussion

As stated in the beginning of this article, films are more frequently used as instructional tools in schools than a variety of other media (Feierabend & Klingler, 2003). However, the effectiveness of films as learning tools has often been questioned because it has been assumed that films are not elaborated (DeFleur et al., 1992; Salomon, 1984; Weidenmann, 1989, 2002). The present article demonstrates that this assumption cannot be supported. Instead, the present three experiments revealed that expository film viewers engage in elaborative processing during expository film comprehension. Film viewers inferred causal information online (Experiment 1), as indicated by shorter naming latencies in an IR context compared with an IU context. Moreover, after watching a video, viewers did not perform worse on questions where a causal connection had to be inferred compared with questions where the relevant causal information was explicitly given. Additionally, at least for one item set, viewers outperformed the control group who did not watch the film on those questions. Both findings indicate that the viewers integrated the inferred causal information into

Table 3
Means (and Standard Deviations) of Comprehension Scores as a Function of Presentation Mode and Coherence of Narration

Presentation mode	Coherence of narration	<i>M</i> (<i>SD</i>)
Video	Coherent	5.67 (0.49)
	Incoherent	5.60 (0.51)
Audio	Coherent	5.47 (0.64)
	Incoherent	4.33 (1.76)

their mental model of the film content (Experiment 2). Additionally, in the first two experiments, film recipients did not differ in their performance from audio recipients. This finding suggests that expository films are elaborated equally well as expository audiotapes and that the dynamic pictorial information does not hinder comprehension. Finally, it was shown that film viewers were able to compensate for a coherence break in the narration by attending to the accompanying pictorial information (Experiment 3). However, this compensation effect was only found on a local level. Overall, narration including a substantial number of coherence breaks led to a decrease in comprehension, irrespective of the existence of pictorial information (Experiment 3).

Taken together, the results from the present experiments question the common assumption that, due to their fast pace, films are not elaborated by their viewers but instead are superficially processed (DeFleur et al., 1992; Furnham et al., 2002; Weidenmann, 1989). In contrast, the present findings provide evidence that viewers engage in online elaboration of the film's content by generating LCB inferences and compensating for coherence breaks. One reason for this seemingly contradictory finding is that previous studies have based their notion of shallow film processing on global comprehension measures attained after watching a film. The present experiments extend these studies in three ways, namely, by measuring comprehension locally, by specifying a particular indicator of elaboration (LCB inferences), and by including an online measure of this indicator. Through this empirical approach, it is shown that, on a local level, film viewers indeed process the film and produce coherent mental representations of its content.

By comparing the results of watching a video that combines narration and pictures with the results of listening to the narration without accompanying pictures, some conclusions can be drawn regarding the role pictorial information plays in comprehension. Both in Experiment 1 and in Experiment 2, no differences between video and audio-only conditions were found with regard to LCB inferences. This raises the question as to what extent film pictures contribute to this process. One possible reason for the lack of differences between the video and the audio-only presentation in the first two experiments could be that the narration was highly coherent. As such, the audio track was sufficient for comprehension. Consistent with this argumentation, Experiment 3 showed that in the case of an incoherent narration, the accompanying pictures, provided via the visual channel, led to better comprehension of the respective parts of the video. Thus, the pictorial information can help viewers to compensate the coherence breaks in educational films or documentaries that are not as coherent and as redundant as educational films aiming at children as the target audience. In line with this argumentation, educational films do indeed show educational potential. They convey information through the pictorial and aural channels, and students can therefore acquire a more comprehensive mental model. Additionally, films show actual objects and realistic scenes, sequences in motion and perspectives that are difficult or impossible to observe in real life (Wetzel et al., 1994).

The beneficial effect of redundant pictures for comprehension of narration corroborates findings from communication studies (Drew & Grimes, 1987; Grimes, 1990; Fox, 2004; Zhou, 2004) and also underscores the differences of video material to animations as investigated in multimedia research (Mayer, 2001, 2005;

Sweller, 1999). Whereas the latter typically consists of simplified graphical depictions that are temporally only loosely coupled to the narration, films typically rely on dense photographic material that is strongly coupled to the narration and that undergoes frequent abrupt changes via film cuts. As a result, a certain amount of redundancy between pictorial and verbal information may both render a message easier to grasp and bolster against coherence breaks in either of the information channels.

Certainly, engaging in online elaboration during viewing, as was observed in the present experiments, does not guarantee that an appropriate mental representation of the film's content is permanently formed. Accordingly, the present findings do not necessarily contradict previous results showing that films may lead to suboptimal comprehension compared with texts (Furnham et al., 2002). In this respect, two results of the present studies are particularly noteworthy. First, in Experiment 2, the differences between the implicit condition and the control group were not unambiguous; they answered the question whether online inferences were firmly integrated into the subsequent mental model. Second, in Experiment 3, the results of the summaries and the near transfer task show that the compensating function of the pictorial information is restricted locally. Both findings suggest that while engaged in online elaboration, the viewers nevertheless were obviously not completely successful in developing an appropriate mental representation of the film's content. A possible explanation for this might be that the process of integrating the results of the elaboration requires additional cognitive resources that compete with the demands of the ongoing film presentation. Here, texts may have certain advantages because they allow the readers to process the information at their own pace, thereby adapting themselves to demands of high-information density (Schwan & Riempp, 2004). In order to achieve a clearer picture of these demands, future studies should not only include measures that are indicative of local variations in cognitive load (such as dual-task measures), but they should also consider the possibilities of controlling video material in an interactive, booklike manner.

References

- Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1061–1070. doi:10.1037/0278-7393.19.5.1061
- Baggett, P. (1979). Structurally equivalent stories in movie and text and the effect of the medium on recall. *Journal of Verbal Learning and Verbal Behavior*, 18, 333–356. doi:10.1016/S0022-5371(79)90191-9
- Boucheix, J.-M., & Lowe, R. K. (2010). An eye tracking comparison of external pointing cues and internal continuous cues in learning with complex animations. *Learning and Instruction*, 20, 123–135. doi:10.1016/j.learninstruc.2009.02.015
- Bovair, S., & Kieras, D. E. (1985). A guide to propositional analysis for research on technical prose. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text* (pp. 315–362). Hillsdale, NJ: Erlbaum.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466. doi:10.1016/S0022-5371(80)90312-6
- DeFleur, M. L., Davenport, L., Cronin, M., & DeFleur, M. (1992). Audience recall of news stories presented by newspaper, computer, television, and radio. *Journalism Quarterly*, 69, 1010–1022. doi:10.1177/107769909206900419

- Drew, D. G., & Grimes, T. (1987). Audiovisual redundancy and TV news recall. *Communication Research*, 14, 452–461. doi:10.1177/009365087014004005
- Feierabend, S., & Klinger, W. (2003). *Kinder und Medien 2002. Ergebnisse der Studie KIM 2002 zum Medienumgang sechs- bis 13-Jähriger in Deutschland* [Results of the KIM 2002 study regarding media exposure of six to 13-year-old children]. Baden-Baden, Germany: Medienpädagogischer Forschungsverbund.
- Fincher-Kiefer, R., & D'Agostino, P. R. (2004). The role of visuospatial resources in generating predictive and bridging inferences. *Discourse Processes*, 37, 205–224. doi:10.1207/s15326950dp3703_2
- Fox, J. R. (2004). A signal detection analysis of audio/video redundancy effects on television news video. *Communication Research*, 31, 524–536.
- Furnham, A., de Siena, S., & Gunter, B. (2002). Children's and adults' recall of children's news stories in both print and audio-visual presentation modalities. *Applied Cognitive Psychology*, 16, 191–210. doi:10.1002/acp.777
- Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 430–445. doi:10.1037/0278-7393.16.3.430
- Gibbons, J., Anderson, D. R., Smith, R., Field, D. E., & Fischer, C. (1986). Young children's recall and reconstruction of audio and audio-visual narratives. *Child Development*, 57, 1014–1023. doi:10.2307/1130375
- Graesser, A. C., & Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading*, 2, 247–269. doi:10.1207/s1532799xssr0203_4
- Graesser, A. C., Léon, J. A., & Otero, J. (2002). Introduction to the psychology of science text comprehension. In J. Otero, A. Léon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 1–15). Mahwah, NJ: Erlbaum.
- Graesser, A., Louwerse, M., McNamara, D., Olney, A., Cai, Z., & Mitchell, H. (2007). Inference generation and cohesion in the construction of situation models: Some connections with computational linguistics. In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 289–310). Mahwah, NJ: Erlbaum.
- Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, 48, 163–189. doi:10.1146/annurev.psych.48.1.163
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371–395. doi:10.1037/0033-295X.101.3.371
- Graesser, A. C., Wiemer-Hastings, P., & Wiemer-Hastings, K. (2001). Constructing inferences and relations during text comprehension. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 249–271). Amsterdam, the Netherlands: John Benjamins.
- Grimes, T. (1990). Audio-visual correspondence and its role in attention and memory. *Educational Technology Research and Development*, 38, 15–25. doi:10.1007/BF02298178
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). Amsterdam, the Netherlands: North-Holland. doi:10.1016/S0166-4115(08)62386-9
- Hayes, D. S., Kelly, S. B., & Mandel, M. (1986). Media differences in children's story synopses: Radio and television contrasted. *Journal of Educational Psychology*, 78, 341–346. doi:10.1037/0022-0663.78.5.341
- Hochberg, J., & Brooks, V. (1978). Film cutting and visual momentum. In J. W. Sanders, D. F. Fisher, & R. A. Monty (Eds.), *Eye movements and the higher psychological functions* (pp. 293–313). Hillsdale, NJ: Erlbaum.
- Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, 20, 146–154. doi:10.1016/j.learninstruc.2009.02.019
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163–182. doi:10.1037/0033-295X.95.2.163
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.
- Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human Computer Studies*, 65, 911–930. doi:10.1016/j.ijhcs.2007.06.005
- Long, D. L., Golding, J. M., & Graesser, A. C. (1992). A test of the on-line status of goal-related inferences. *Journal of Memory and Language*, 31, 634–647. doi:10.1016/0749-596X(92)90032-S
- Lowe, R. K. (2003). Animation and learning: Selective processing of information in dynamic graphics. *Learning and Instruction*, 13, 157–176. doi:10.1016/S0959-4752(02)00018-X
- Lowe, R. K. (2004). Interrogation of a dynamic visualization during learning. *Learning and Instruction*, 14, 257–274. doi:10.1016/j.learninstruc.2004.06.003
- Magliano, J. P., Dijkstra, K., & Zwaan, R. A. (1996). Generating predictive inferences while viewing a movie. *Discourse Processes*, 22, 199–224. doi:10.1080/01638539609544973
- Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology*, 15, 533–545. doi:10.1002/acp.724
- Maiwald, A. (Writer), & Biermann, C. (Director). (2006a). *Abblendenspiegel* [Rear view mirror] [Television series episode]. In J. Lachmuth (Producer), *Sachgeschichten aus der Sendung mit der Maus*. Köln, Germany: WDR.
- Maiwald, A. (Writer), & Biermann, C. (Director). (2006b). *Blitzentstehung* [Development of a lightning] [Television series episode]. In J. Lachmuth (Producer), *Sachgeschichten aus der Sendung mit der Maus*. Köln, Germany: WDR.
- Maiwald, A. (Writer), & Biermann, C. (Director). (2006c). *Thermoskanne* [Thermos] [Television series episode]. In J. Lachmuth (Producer), *Sachgeschichten aus der Sendung mit der Maus*. Köln, Germany: WDR.
- Mayer, R. E. (2001). *Multimedia learning*. Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9781139164603
- Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9780511816819.004
- Millis, K. K., & Graesser, A. C. (1994). The time-course of constructing knowledge-based inferences for scientific texts. *Journal of Memory and Language*, 33, 583–599. doi:10.1006/jmla.1994.1028
- Murray, J. D., & Burke, K. A. (2003). Activation and encoding of predictive inferences: The role of reading skill. *Discourse Processes*, 35, 81–102. doi:10.1207/S15326950DP3502_1
- Noordman, L. G. M., Vonk, W., & Kempff, H. J. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language*, 31, 573–590. doi:10.1016/0749-596X(92)90029-W
- Potts, G. R., Keenan, J. M., & Golding, J. M. (1988). Assessing the occurrence of elaborative inferences: Lexical decision versus naming. *Journal of Memory and Language*, 27, 399–415. doi:10.1016/0749-596X(88)90064-2
- Robinson, G. J. (1988). *Emotional effects of media: The work of Hertha Sturm*. Montreal, Canada: McGill University.
- Salomon, G. (1984). Television is “easy” and print is “tough”: The differential investment of mental effort in learning as a function of perceptions and attribution. *Journal of Educational Psychology*, 76, 647–658. doi:10.1037/0022-0663.76.4.647

- Schwan, S., & Riempp, R. (2004). The cognitive benefits of interactive videos: Learning to tie nautical knots. *Learning and Instruction, 14*, 293–305. doi:10.1016/j.learninstruc.2004.06.005
- Singer, M., Harkness, D., & Stewart, S. T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes, 24*, 199–228. doi:10.1080/01638539709545013
- Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER.
- Tversky, B., Bauer-Morrison, J., & Bétrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies, 57*, 247–262.
- Walma van der Molen, J. H., & Van der Voort, T. H. A. (2000). Children's and adults' recall of television and print news in children's and adult news formats *Communication Research, 27*, 132–160.
- Weidenmann, B. (1989). Der mentale Aufwand beim Fernsehen [The mental effort during watching TV]. In J. Groebel & P. Winterhoff-Spurk (Eds.), *Empirische Medienpsychologie* (pp. 134–149). München, Germany: Psychologie-Verlags-Union
- Weidenmann, B. (2002). Abbilder in Multimedia-Anwendungen [Representations in multimedia applications]. In L. J. Issing & P. Klimsa (Eds.), *Information und Lernen mit Multimedia* (pp. 83–96). Weinheim, Germany: Psychologie Verlags Union.
- Wetzel, C. D., Radtke, P. H., & Stern, H. W. (1994). *Instructional effectiveness of video media*. Hillsday, NJ: Erlbaum.
- Wiley, J., & Myers, J. L. (2003). Availability and accessibility of information and causal inferences from scientific text. *Discourse Processes, 26*, 109–129.
- Zhou, S. (2004). Effects of visual intensity and audiovisual redundancy in bad news. *Media Psychology, 6*, 237–256. doi:10.1207/s1532785xmep0603_1
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.

Received February 27, 2009

Revision received August 6, 2012

Accepted October 11, 2012 ■

Managing Face Threats and Instructions in Online Tutoring

Benjamin Brummernhenrich and Regina Jucks
Westfälische Wilhelms-Universität Münster

Although tutoring is very effective, tutors often neglect certain strategies such as direct negative feedback. This might be because they want to avoid threatening their tutee's face. The concept of face derives from politeness theory and refers to the aspects of autonomy and social appreciation people claim for themselves and strive to negotiate cooperatively in discourse. We argue that tutors' politeness considerations can hinder effective tutoring. We compared 2 interventions in naturalistic tutoring interactions designed to influence tutors' communication acts: In a politeness condition, tutors were advised not to restrict autonomy when explaining concepts or correcting them in order to save their tutee's face. In a no-politeness condition, in contrast, they were encouraged to communicate clearly and explicitly. Results showed that tutors in the no-politeness condition used not only more direct strategies such as requests and hints, but, unexpectedly, also more politeness strategies to mitigate their directness. We conclude that there is a clear connection between politeness and instructional moves, but it remains unclear whether tutors avoid instructional face threats because they construe them as face-threatening. We suggest how further research could cast light on the conditions under which politeness is detrimental or beneficial for tutoring.

Keywords: human tutoring, politeness theory, face threats, instructional communication

Tutoring is a very effective form of instruction (Cohen, Kulik, & Kulik, 1982; Graesser, D'Mello, & Cade, 2011), and over the past 20 years, research has identified typical tutoring behaviors and pinpointed those that effectively help tutees achieve their tasks (Graesser, Person, & Magliano, 1995; Lepper & Woolverton, 2002). However, there are some tutorial moves that tutors seem reluctant to use. For example, tutors are much slower in giving feedback on error-ridden tutee statements than on correct ones (Cromley & Azevedo, 2005) and hardly ever correct tutees' mistakes explicitly (Chi, Siler, & Jeong, 2004; Lepper & Woolverton, 2002). Attending to misconceptions is, however, vital in instructional explanations (Wittwer & Renkl, 2008). One explanation for this behavior might be tutors' inability to monitor their tutees' understanding. In contrast, Person, Kreuz, Zwaan, and Graesser (1995) have argued that tutors' reluctance to use certain instructional strategies is due to their desire not to impose on their tutees. They suppose that certain strategies such as negative feedback or requesting actions are incompatible with the politeness principles to which tutors try to adhere.

Politeness in Instruction

In their seminal work on politeness, Brown and Levinson (1987) posited that certain speech acts threaten the *face* of the interlocutor, defined as the "positive social value a person effectively claims for himself" (Goffman, 1967, p. 5). Many of the cognitive strategies commonly applied by tutors (D'Mello, Olney, & Person, 2010) arguably necessitate face-threatening acts (FTAs). Following Brown and Levinson's (1987) terminology, negative feedback such as "no, that's wrong" threatens the tutees' *positive face*, the desire for social acceptance and belonging. *Negative face*, the desire to be autonomous and unrestricted in one's actions, can, for example, be threatened by prompts such as "Tell me what you know about that topic" and posing a new task such as "You have to calculate the effect size now," because these demand a specific action or an answer. When a tutorial dialogue reaches a point at which a tutor would have to perform an FTA to move the tutoring forward, tutors have three options: not perform the FTA, do it baldly, or "redress" it using positive or negative *politeness strategies* (Brown & Levinson, 1987). Positive politeness strategies aim to reduce the threat to the hearer's face by emphasizing common interests and by assuring the hearer that her or his needs are respected. Negative politeness strategies try to reduce the face threat itself by minimizing the imposition and granting autonomy to the hearer.

Most research has centered on positive effects of politeness: Polite e-mail is evaluated more positively than e-mail with unmitigated face threats, and its senders are perceived as friendlier, more likable, and more competent (Jessmer & Anderson, 2001). Unmitigated face threats, on the other hand, can evoke negative emotions and feelings of unfairness (Carson & Cupach, 2000; Cupach & Carson, 2002). Correspondingly, tutors who apply facework are perceived as being more credible, likable, and recipient oriented than those who apply less facework (Jucks, Brummernhenrich, &

This article was published Online First February 18, 2013.

Benjamin Brummernhenrich and Regina Jucks, Department of Psychology and Sport Studies, Westfälische Wilhelms-Universität Münster, Münster, Germany.

We are grateful to Arthur Graesser for valuable comments on a draft of this article. We also wish to thank Frieda Lina Bense, Gesa Linnemann, and Lena Päuler for their help with data collection and analysis, as well as Jonathan Harrow for language editing.

Correspondence concerning this article should be addressed to Benjamin Brummernhenrich, Institute of Psychology for Education, Westfälische Wilhelms-Universität Münster, Fliegerstrasse 21, 48149 Münster, Germany. E-mail: brummernhenrich@uni-muenster.de

Päuler, 2012). Polite computer tutors also achieve higher learning gains in their tutees than nonpolite tutors (Wang et al., 2008). This might be due to higher tutee motivation and involvement (Kerssen-Griep, Hess, & Trees, 2003), which, in turn, lead to better knowledge construction.

Other research suggests that being polite hinders learning. Person et al. (1995), for example, performed detailed analyses of tutoring transcripts showing how politeness considerations led tutors to miss opportunities for effective instruction and thus deliver suboptimal tutoring. Apart from the more obvious drawbacks of not giving the tutee explicit feedback, polite redress of instructional FTAs may also make the tutor's discourse less understandable for the tutee. According to politeness theory, negative politeness strategies aim to reduce the illocutionary force of an utterance (Brown & Levinson, 1987). Indeed, there is empirical evidence from other contexts that polite formulations make the specific intent of utterances ambiguous (Bonnefon, Feeney, & De Neys, 2011; Kallia, 2005). Patients, for example, tend to be unsure whether they should understand the use of the hedge *possibly* in sentences like "You will possibly suffer from deafness soon" as a tactful mitigation of a potentially face-threatening utterance or as a sign of real uncertainty (Pighin & Bonnefon, 2011). If tutors use polite formulations because they perceive certain instructional strategies to be face-threatening, this effect could have detrimental consequences in learning contexts. There is evidence that this is the case; medical experts responding to a layperson's error-ridden request for information explained more and wrote longer texts when explicitly instructed to disregard politeness (Bromme, Brummernhenrich, Becker, & Jucks, 2012).

Rationale for the Present Study

Our study focuses on politeness in tutors' communication. Following Person et al. (1995), we assume that tutors avoid or redress certain tutorial strategies because they construe them as face-threatening. If this is the case, "switching off" tutors' politeness should increase the proportion of tutorial FTAs. This, in turn, should improve the tutors' instructional communication and, ultimately, lead to better learning outcomes. The present study pursues this idea by instructing some tutors to communicate to tutees clearly and explicitly and comparing these tutors with others who are encouraged specifically to use negative politeness. Negative politeness strategies reduce the face threat itself by minimizing the imposition and thereby reducing illocutionary force. Therefore, these strategies should have a more detrimental effect on the clarity of tutors' utterances than positive strategies that do not change the speech act itself (Brown & Levinson, 1987).

We expect that tutors who are instructed to be clear and explicit will redress fewer FTAs and use more cognitive scaffolding, give more negative feedback, and provide more elaborate explanations than those who are instructed to use negatively polite communication. If politeness has the suspected effects on tutoring, this instruction should also impact on the tutees' reactions to their tutors' communication, that is, on the way they perceive their tutor or their task motivation. As research on the effects of politeness on tutoring outcomes is inconsistent, we do not state directional hypotheses for these variables. Nonetheless, it is likely that polite tutors will be rated more positively on variables pertaining to social perceptions.

Method

We set up a naturalistic tutoring situation. Advanced psychology students tutored novice students on how to perform an analysis of variance (ANOVA) with the Statistical Package for Social Sciences (SPSS). The tutors were sufficiently adept at this task, whereas the tutees had little experience with using SPSS. Tutors and tutees were located at different computers and communicated via an Instant Messenger chat program.

Design and Materials

Manipulation. In two experimental conditions, tutors received different instructions before the tutoring session. Whereas both instructions stated that tutees might feel offended by explanations, requests, or corrections, they differed in the specific reasons given for this offense. The instructions ended by giving tutors three hints on how to avoid these situations. These hints were condition specific and aimed to prompt different politeness strategies.

In the politeness condition, we informed tutors that tutees might feel offended by having their autonomy restricted. We instructed tutors with three hints prompting them not to pressure tutees and to phrase explanations and corrections in a way that would leave the tutees a choice regarding how to respond to them. An example of a hint in the politeness condition is "Phrase your *requests noncoercively*! Tutees should feel that they are free to choose whether to respond to your suggestions."

In the no-politeness condition, we informed tutors that tutees might feel offended by not understanding what they have to do. We instructed tutors to communicate clearly and explicitly and gave them no information on the role of politeness. An example of a hint in this condition is "Phrase *requests clearly and explicitly*! Tutees should know exactly what they need to do."

The three hints were also shown to the tutors as prompts during the tutoring session itself. During the first 30 min of the session, a prompt was shown every 5 min (so that each hint appeared twice during the session). This resulted in a maximum of six hints, given that the session did not last less than 30 min (see the Appendix for English translations of the instructions and all prompts). Figure 1 provides a screenshot of the online environment. We performed no manipulations with the tutees.

Tutoring task. Tutees were given two tasks: one during the tutoring session and one to be completed alone after the tutoring. Although the two tasks used different data sets from real psychological studies, they required the same standard statistical calculations taught routinely in psychological research methods courses and textbooks (e.g., Howitt & Cramer, 2008). The tutees performed the ANOVA with SPSS and reported their interpretations with text processing software.

Participants

Participants were 68 university students from a large German university with at least some background in data analysis. They formed 34 tutor-tutee dyads. All tutors were at least in their third year of psychology studies, had completed advanced statistics courses, and had used SPSS on several occasions. Of the tutees, 30 were in their second year; the remaining four were students of

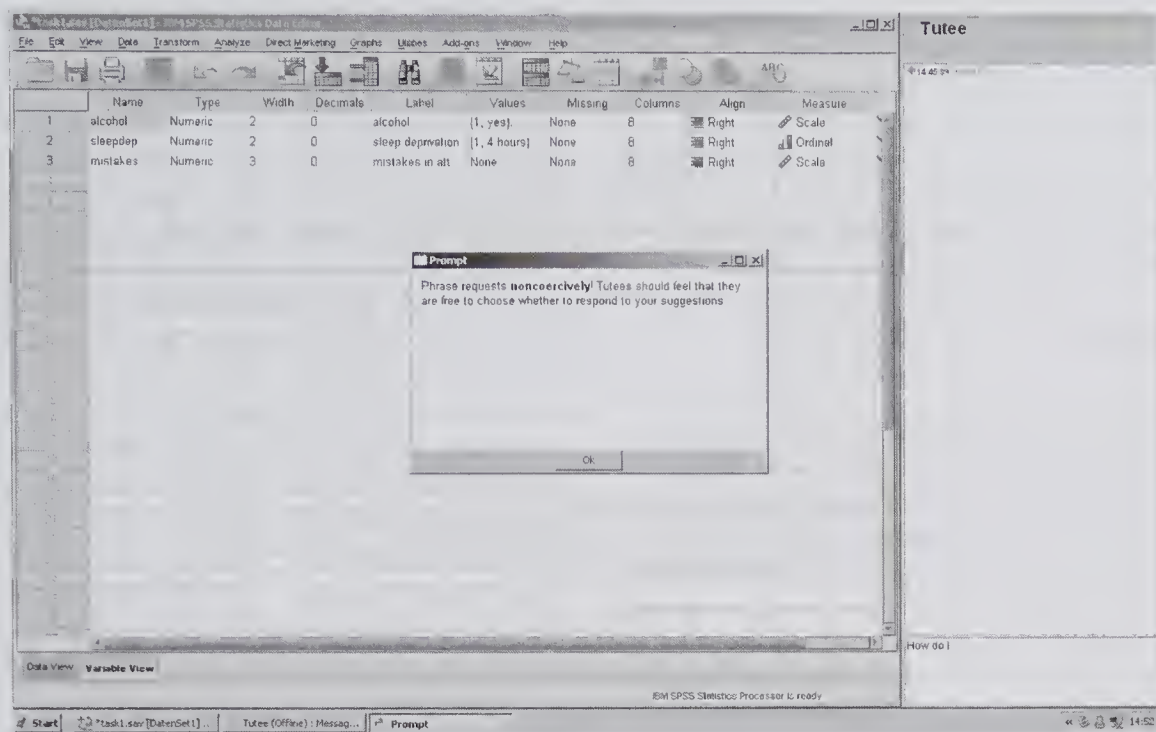


Figure 1. Screenshot of the online environment: Tutee's SPSS window on the left-hand side, chat window on the right, tutor's prompt in the center.

other subjects (e.g., sociology, politics) who had completed only a basic statistics course covering roughly the same material as that taught to novice psychology students. Participants were recruited in the corresponding statistics lectures and courses or on the university's psychology campus and received either €10 (about \$13) or course credits. The tutor-tutee dyads were assigned randomly to one of the two experimental conditions.

Expertise and competence of tutors. The tutors performed the ANOVA task before the tutoring session, and we evaluated the correctness of their solution as a measure of their task expertise and competence. We had to exclude the data of three dyads in which the tutors solved less than one-half of the steps correctly. Both tutors and tutees were asked to rate their understanding of several task-relevant statistical concepts (*t* tests, reason for using ANOVA, interpretation of interactions, meaning of Q-Q plots, effect sizes) and report the number of different contexts in which they had used SPSS. The remaining tutors scored significantly higher on these measures than their respective tutees (using Wilcoxon signed-rank tests due to nonnormality; concepts: $W = 115.5$, $p = .047$; contexts: $W = 54$, $p = .030$).

Final sample composition. The data set consists of 31 tutor-tutee dyads: 15 in the politeness condition and 16 in the no-politeness condition. Twenty-eight (90%) of the tutors and 24 (77%) of the tutees were female. The high proportion of female participants reflects the fact that the population of psychology students from which we drew our sample was itself roughly 80% female. The mean ages of tutors and tutees were 24.48 years ($SD = 4.72$) and 23.26 years ($SD = 3.40$), respectively. Preliminary analyses revealed no significant differences between experimental conditions in terms of age; gender ratio; weekly computer use; frequency of chat, e-mail, and Internet forum use; self-rated knowledge of statistical concepts; and tutors' solutions of the ANOVA transfer task. However, marginally significant differences between conditions were found for tutors' weekly Internet

use ($W = 168.5$, $p = .057$), tutors' frequency of messenger program use ($\chi^2 = 8.93$, $p = .063$), and tutors' experience with SPSS ($W = 167.5$, $p = .056$). The tutees' data did not show these differences. Therefore, we entered these variables as covariates in all analyses of the tutors' data.

Procedure

Tutees and tutors were greeted by separate experimenters and led to different rooms. They each worked on laptops with 15-in. (37.10-cm) screens displaying all materials and questions except the ANOVA task descriptions, which were handed to the participants on paper.

Individual phase. At the start of the experiment, each participant completed the questionnaire containing the control variables described above. Then the tutors completed the ANOVA task once by themselves. If they were unsure about a specific step, the experimenter first encouraged them to try it by themselves. If the participant repeated the request, the experimenter told them only the correct menu item in SPSS, but did not instruct them any further so as not to prime any specific phrasing or instructional strategy that tutors might adopt during the subsequent tutoring session.

Tutoring session. After this, both tutors and tutees received their instructions for the tutoring sessions and the tutees received the task. The chat program was started on both computers, appearing as a narrow window on the right-hand side of the screens. SPSS and the text processing software were started on the tutee's laptop. These windows and the tutee's actions were also visible on the tutor's screen. When both tutor and tutee signaled that they were ready, a test message was posted in the chat window and the tutoring session started. The chat log and screen content were recorded for later analysis. The tutoring ended when the tutee finished the task.

Posttutoring questionnaires. Tutors now completed the post-tutoring questionnaire marking the end of the experiment for them. The tutees completed the first part of the posttutoring questionnaire containing questions on the tutoring itself and then performed the second task during which the screen content was again recorded. After this, they completed the second part of the post-tutoring questionnaire. The whole session took about 90 min for both tutors and tutees.

Measures and Analysis

Data from tutors.

Content analysis of tutor communication. Our foremost interest was in how the different instructions impacted on the tutors' communication. We used only the discourse data of the last subtask, because this required both declarative and procedural components of expertise about statistics and SPSS. Tutees were asked to write down their interpretation of the results of the several statistical analyses they had performed. During this step, tutors frequently guided their tutees in reviewing the output from SPSS and the task description, so that they would arrive at the correct conclusions. The chat logs of this subtask were subjected to a content analysis (Chi, 1997). We chose noncontent features as boundaries for segmenting the discourse data: message endings, commas, or conjunctions such as *and*, *because*, and so forth. Two trained coders who were blind to the experimental conditions coded the resulting segments on the following variables:

Face-threatening acts. The categories for the content analysis of the tutor's use of FTAs were derived directly from Brown and Levinson's (1987) theory: As depicted in Figure 2, the rater first decided whether an utterance constituted an FTA. If this was the case, the rater decided whether it was expressed baldly (without any redress), whether it was off record, or whether it was redressed with positive or negative politeness strategies. In cases where a single utterance was redressed by both types of politeness strategies, we gave precedence to negative politeness, because this was the focus of our study.

Instructional behaviors. We then categorized which kinds of instructional behaviors the tutors displayed in a given segment using D'Mello, Olney, and Person's (2010) list of 27 categories of tutor moves. Because this list had been compiled by analyzing tutoring interactions on declarative knowledge, we adapted it to the mixed declarative/procedural context of our study by adding categories such as requesting the tutee to perform an action. We also dropped categories that confound tutorial objectives with politeness considerations (e.g., neutral feedback and motivational statement). Trial runs with initial codings led to further adaptations of

the categorization system, resulting in a final 14 categories. To increase the clarity of the results, we subsumed nearly all moves that made up less than 3% of utterances in both conditions into an other-instructional-behaviors category, with one exception: negative feedback, because this was especially interesting from a theoretical viewpoint (VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003; Wittwer & Renkl, 2008). The final set of analyzed tutoring moves contained nine strategies: explanation, request for action, hint, give correct answer, clarification question, positive feedback, negative feedback, scaffolding (prompts and pumps), and other instructional behaviors.

Examples for all categories of FTAs and tutoring behaviors are presented in Table 1. Four of the sessions were coded by both raters to assess interrater agreement. The interrater reliability was acceptable for both politeness moves (Krippendorff's $\alpha = .79$) and instructional moves ($\alpha = .70$).

Self-report measures. The following measures were assessed on a 5-point scale ranging from 1 (*low*) to 5 (*high*) indicating the participant's agreement with each of the statements.

Evaluation of prompts. Tutors reported whether they had been able to purposefully apply the communication strategies explained to them in the instructions and the prompts, and to evaluate what use they had been. The items were "The instructions were helpful," "I was able to apply the instructions," "It's a good idea to phrase explanations, requests, and corrections in this way," and "The messages bothered me during the communication."

Tutor's perception of tutee's understanding. The tutors rated how they perceived their tutee's understanding of the topic on five items adapted from the Recipient Orientation Scale (ROS; Bromme, Jucks, & Runde, 2005; see below). A sample item is "The tutee would now be able to explain what you have taught to a friend." The reliability of this measure was acceptable (Cronbach's $\alpha = .87$).

Data from tutees.

ANOVA transfer task. To evaluate the learning outcomes of tutoring, we assessed the tutees' competence on a second ANOVA task. We defined 11 partial solutions and gave the tutees 1 point for each one that was correct. These solutions consisted of either certain SPSS outputs (e.g., the correct diagrams that had been asked for in the task) or propositions that should appear in the written interpretations (e.g., "The interaction is not significant"). Thus, every tutee received a score between 0 and 11 points.

Self-report measures. As with the tutors, all tutees' self-reports were measured on 5-point scales.

Knowledge of concepts and attitude toward statistics. To assess differences between experimental conditions in the tutees'

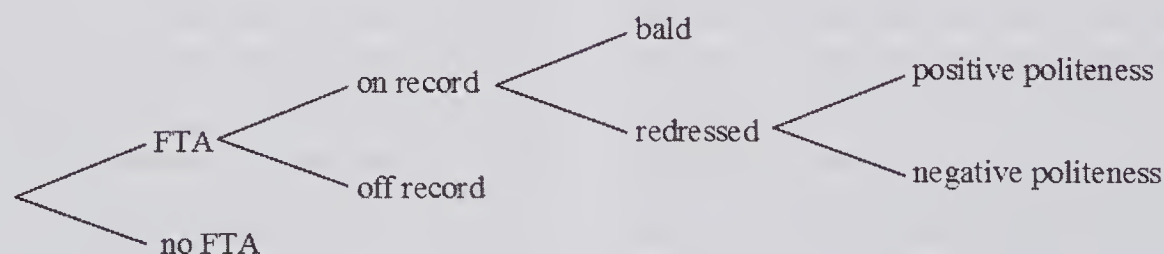


Figure 2. Categories for content analysis of face-threatening acts (FTAs). Adapted from *Politeness: Some Universals in Language Usage* (p. 69), by P. Brown and S. C. Levinson, 1987, Cambridge, England: Cambridge University Press. Copyright 1987 by Cambridge University Press.

Table 1
Examples of Face-Threatening Acts and Instructional Behaviors

Tutor behavior	Example utterance
Face-threatening acts	
Bald/on record	Now look whether that value is larger than .14.
Positive politeness	The first thing we want to know is whether the main effect is significant.
Negative politeness	Why don't you write that down as the interpretation?
Off record	The effect is also visible in the other conditions of that independent variable. (Adding an aspect the tutee has neglected to include in her written interpretation)
Instructional behaviors	
Explanation	Eta square is a measure of effect size.
Request for action	Scroll down now.
Hint	The plots can also show you whether they might be an interaction.
Give correct answer	The interaction is not significant.
Clarification question	Are you looking for something specific?
Positive feedback	Yes, that's right.
Negative feedback	No, those aren't categories.
Scaffolding (prompts and pumps)	What else could you choose apart from an "interval scale"?
Other instructional behaviors	Let's turn to the second problem now.

subjective learning gain, we measured the perceived understanding of five concepts central to the task described in the Participants section (five items; pretest Cronbach's $\alpha = .81$) in a pre- and posttest design. We also used a pre- and posttest design to ask the tutees to indicate their attitude toward three aspects of using statistics: "I am interested in statistics," "I feel confident when using SPSS," and "I feel competent regarding statistics" (pretest Cronbach's $\alpha = .82$).

Tutor's recipient orientation. The ROS (Bromme et al., 2005) was used to measure how the tutees perceived four aspects of their tutors' communication: (a) audience design, that is, the extent to which the tutee gained the impression that the tutor tried to take a layperson's perspective into account (e.g., "My tutor knows exactly what my problems are with this topic"; 10 items); (b) subjective appraisal of own comprehension (e.g., "I could explain the things I have been taught to a friend"; five items); (c) tutor's specialized knowledge and commitment to writing on this specific issue (e.g., "My tutor is very knowledgeable in this field"; seven items); and (d) emotional evaluation of the tutoring (e.g., "I enjoyed reading the tutor's explanations"; nine items). All subscales had satisfactory consistencies with a Cronbach's alpha of at least .64 (for perceived audience design and .78–.87 for the other three subscales).

Tutor's instructional facework. We used the Revised Instructional Face-Support Scale (Kerssen-Griep, Trees, & Hess, 2008) to measure how tutees perceived the use of positive (four items) and negative politeness (four items) strategies by their tutors. Sample items are "My tutor made sure that she or he didn't cast me in a bad light" (positive politeness) and "My tutor left me free to choose how to respond" (negative politeness). Scale consistencies were satisfactory with alpha values of .76 and .60, respectively.

Intrinsic motivation. We used the intrinsic motivation subscale taken from the Situational Motivation Scale (Guay, Vallerand, & Blanchard, 2000; four items; Cronbach's $\alpha = .88$) to measure the tutees' intrinsic motivation to perform the ANOVA task. A sample item is "I found working on this task interesting."

Social relationship to tutor. We assessed the appraisal of the social relationship to the tutor by adapting the following items

from a scale measuring relational trust (Echterhoff, Lang, Krämer, & Higgins, 2009): "I find my tutor likeable," "I would work with my tutor again," "I feel close to my tutor," "I feel connected to my tutor through our communication," and "I find it important to see the problem in the same way as my tutor." Because these items do not form a homogeneous scale, they were analyzed individually.

Control variables. We entered the pretest measures of the tutees' perceived understanding of concepts central to the task and the attitude toward statistics and SPSS analyses as covariates.

Results

Preliminary Analyses

Preliminary analyses revealed that most data suitable for common ANOVA-like procedures were nonnormal and some were heteroscedastic. Therefore, Winsorized variances were used to provide more robust estimators (Erceg-Hurn & Mirosevich, 2008). The data from the content analysis were handled differently, because they consisted of proportions and not continuous variables (see below).

As reported earlier, we entered the variables weekly Internet use, frequency of messenger program use, and experience with SPSS as covariates in all analyses of tutors' data, and all descriptive values reported were adjusted for the mean level of these variables.

Data From Tutors

Content analysis. To analyze whether the proportions of different instructional behaviors and politeness strategies differed between experimental conditions, we divided the number of segments showing a certain tutor behavior (e.g., number of segments with hints, number of segments with positive politeness) by the number of total segments in the interaction. This yielded the proportions of different instructional behaviors and politeness strategies for each tutor. This type of data (compositional data) cannot be modeled by common procedures such as logistic regres-

sion, because there are more than two response categories and more than one response in every case. Therefore, we used Dirichlet regression models (Gueorguieva, Rosenheck, & Zelterman, 2008; Hijazi & Jernigan, 2009).

Face-threatening acts. Table 2 reports the mean proportions of FTAs and politeness strategies and their standard deviations in each of the conditions along with values for *z* and *p* as well as effect sizes. About one-third of the utterances (31.58%) constituted face threats. Most FTAs were redressed with negative politeness (54.63%). The regression analysis showed that tutors in the no-politeness condition uttered significantly more FTAs baldly (*z* = 2.00, *p* = .045, *d* = 0.75; medium effect) but also used significantly more negative politeness strategies (*z* = 2.29, *p* = .022, *d* = 0.87; large effect). They showed a tendency toward using fewer positively redressed FTAs (*z* = 1.78, *p* = .075, *d* = 0.66; medium effect). Off-record strategies were used very rarely (3.55%), and there was no significant difference between conditions (*z* = 0.369, *p* = .712).

Instructional behaviors. The results for instructional behaviors are shown in Table 3. Of all utterances that constituted instructional moves, hinting was the most prevalent. In the politeness condition, more than one-third of tutors' utterances were hints. This was significantly lower than the proportion of hinting in the no-politeness condition (over 40%; *z* = 1.98, *p* = .048, *d* = 0.74; medium effect). Tutors in the no-politeness condition also requested more actions (*z* = 3.20, *p* = .001, *d* = 1.27; large effect) and gave more correct answers (*z* = 3.28, *p* = .001, *d* = 1.31; large effect) than those in the politeness condition. Other common instructional moves for which we found no significant differences were positive feedback, which made up about one-quarter of tutors' utterances in both conditions, and, to a lesser extent, scaffolding (6.83%).

Self-report measures.
Evaluation of prompts. The tutors' ratings of the instructions and prompts showed no differences between experimental conditions. The tutors found them, on average, not very helpful (*M* = 2.45, *SD* = 1.21), *t*(26) = 1.58 *p* = .558, but they also indicated that they did not find them distracting while tutoring (*M* = 3.41, *SD* = 1.31), *t*(26) = -0.20 *p* = .843. They thought it was sensible to communicate in the manner indicated by the prompts (*M* = 3.55, *SD* = 1.15), *t*(26) = -0.98, *p* = .338, and also generally agreed that they were able to implement the instructions (*M* = 3.42, *SD* = 1.31), *t*(26) = 0.37, *p* = .714.

Tutor's perception of tutee's understanding. There was no significant difference between groups regarding how tutors eval-

uated their tutees' subjective understanding, *t*(26) = 1.53, *p* = .139. The tutors thought their tutees had generally understood the topic (*M* = 3.51, *SD* = 0.72).

Data From Tutees

ANOVA transfer task. Tutees in the politeness condition scored an average of 8.63 (*SD* = 1.86) correct points on the second ANOVA task compared to an average score of 8.13 (*SD* = 2.47) for those in the no-politeness condition. This difference did not attain statistical significance after controlling for tutees' perceived understanding of concepts central to the task and their attitude toward statistics and SPSS, *t*(27) = 0.35, *p* = .727. However, the covariate measuring personal attitude toward statistics was significant, *t*(27) = 2.11, *p* = .044, *d* = 0.76 (medium effect), with more positive attitudes predicting better outcomes on the task.

Self-report measures.
Knowledge of concepts and attitude toward statistics. Before the tutoring session, tutees rated their understanding of five central statistical concepts at 3.66 (*SD* = 0.83); after the session, the estimate was 4.00 (*SD* = 0.42). The pre- and posttest difference was significant, *t*(30) = 3.59, *p* = .001, *d* = 1.29 (large effect), but the gains did not differ between experimental conditions, *t*(27) = -0.75, *p* = .458. The same held for the tutees' attitude toward statistics and SPSS (pretest: *M* = 2.77, *SD* = 0.75; posttest: *M* = 2.90, *SD* = 0.84): significant pre- and posttest difference, *t*(30) = 4.03, *p* < .001, *d* = 1.45 (large effect); difference between conditions, *t*(27) = -0.35, *p* = .733.

Tutor's recipient orientation. There were no significant differences between experimental conditions for the ROS scales for audience design, *t*(27) = 0.485, *p* = .631; tutor's specialized knowledge and commitment, *t*(27) = -0.08, *p* = .938; or emotional evaluation, *t*(27) = 0.84, *p* = .408. However, tutees in the no-politeness condition indicated a marginally higher subjective understanding (*M* = 4.09, *SD* = 0.45) than those in the politeness condition (*M* = 3.69, *SD* = 0.73), *t*(27) = 1.72, *p* = .098, *d* = 0.62 (medium effect).

Tutor's instructional facework. Tutees in the two experimental conditions did not differ in how they perceived their tutors' facework. Facework was assessed as high for both positive (*M* = 4.28, *SD* = 0.59), *t*(27) = 0.84, *p* = .409, and negative (*M* = 4.02, *SD* = 0.61) dimensions, *t*(27) = -0.55, *p* = .590.

Intrinsic motivation. Tutees generally stated that they were intrinsically motivated to do the task (*M* = 3.78, *SD* = 0.76). This

Table 2
Relative Frequencies of Face-Threatening Acts (Adjusted for Covariates)

Face-threatening act	Experimental condition				<i>z</i>	<i>p</i>	Cohen's <i>d</i>
	Politeness		No-politeness				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Bald/on record	.273	.061	.293	.038	2.00	.045*	0.75
Positive politeness	.128	.014	.117	.008	1.78	.075 [†]	0.66
Negative politeness	.516	.066	.548	.038	2.29	.022*	0.87
Off record	.083	.023	.042	.009	0.37	.712	

† *p* < .10. * *p* < .05.

Table 3
Relative Frequencies of Instructional Behaviors (Adjusted for Covariates)

Instructional behavior	Experimental condition				<i>z</i>	<i>p</i>	Cohen's <i>d</i>
	Politeness		No-politeness				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Explanation	.081	.024	.062	.015	0.07	.943	
Request for action	.042	.011	.076	.015	3.20	.001**	1.27
Hint	.350	.016	.430	.011	1.98	.048*	0.74
Give correct answer	.029	.005	.064	.007	3.28	.001**	1.31
Clarification question	.059	.011	.027	.002	−1.07	.284	
Positive feedback	.282	.008	.240	.006	1.06	.288	
Negative feedback	.029	.003	.022	.002	0.60	.547	
Scaffolding (prompts and pumps)	.102	.031	.056	.012	0.20	.842	
Other instructional behaviors	.026	.004	.024	.002	1.36	.174	

* $p < .05$. ** $p < .01$.

did not differ between the experimental conditions, $t(27) = 1.56$, $p = .131$.

Social relationship to the tutor. Tutees' perceived social relationship to their respective tutors did not differ between conditions for the questions asking for liking of the tutor, $t(27) = -0.12$, $p = .908$; feeling connected to the tutor, $t(27) = -0.30$, $p = .768$; or finding it important to see the problem in the same way as the tutor, $t(27) = 1.34$, $p = .192$. The analysis revealed two marginally significant results: Tutees in the no-politeness condition felt closer to their tutor ($M = 2.87$, $SD = 1.19$) than those in the politeness condition ($M = 2.00$, $SD = 0.73$), $t(27) = 1.80$, $p = .082$, $d = 0.65$ (medium effect), and were also more in favor of working with their tutor again (no-politeness: $M = 4.40$, $SD = 0.83$, politeness: $M = 3.34$, $SD = 0.68$), $t(27) = 1.84$, $p = .076$, $d = 0.66$ (medium effect).

Discussion

This study examined whether tutors can be instructed to adjust their communication to a specific kind of politeness and whether this impacts on their use of certain instructional strategies. In a computer-mediated tutoring situation, tutors were presented with either instructions or prompts detailing the use of negative politeness strategies with the aim of preserving the tutees' autonomy, or they were instructed to communicate clearly and explicitly.

In part, this manipulation had the desired effect: Whereas tutors in the no-politeness condition did not utter more FTAs, they more frequently uttered them baldly than tutors in the politeness condition. Thus, our data provide evidence that it is possible to influence the politeness of tutors' communication. However, instructing tutors to disregard politeness considerations entirely seems much more difficult, as the counterintuitive results on politeness strategies show: Tutors in the no-politeness condition redressed significantly more FTAs with negative politeness strategies and showed a tendency to use fewer positive strategies. Our concern for the other's face seems to be intertwined deeply with the processes that govern our communication, and it cannot just be switched off by an experimental manipulation.

The data on instructional behaviors were inconsistent: Tutors in the no-politeness condition used more requests and hints and gave more correct answers. These are arguably more direct forms of

instruction, and more likely to constitute threats to the tutee's negative face, explaining why these tutors might have been tempted to make up for their brashness by using negative politeness strategies. Contrary to our expectations, there were no differences for the other instructional strategies such as explanations and scaffolding moves. It is possible that the tutors did not construe these moves as face-threatening and as such they were not influenced by the instructions. However, because we did not investigate how the tutors perceived the appropriateness of their communication, we cannot validate this assumption directly. The fact that tutors in the no-politeness condition used not only more direct instructional moves but also more politeness strategies could indicate that they were aware of the face threat these moves present. On the one hand, this supports Person et al.'s (1995) assumption that politeness considerations hinder the use of face-threatening tutoring moves. On the other hand, most instructional strategies such as scaffolding questions and, most importantly, negative feedback on errors were not influenced by the instruction.

The proportion of negative feedback was very low in this study (just over 1% of all utterances), especially when compared to the high proportion of positive feedback. This was not due to a lack of opportunities, because tutees' errors during the tutoring were frequent. Whereas other studies have found that (expert) tutors' feedback can be very prompt and discriminating (D'Mello, Lehman, & Person, 2010), it is highly unlikely that our tutors gave feedback on every tutee mistake. These results are in line with other research on tutor behavior (Chi et al., 2004; Lepper & Woolverton, 2002) and could be due to tutors' awareness of the face threat inherent in negative feedback, inducing them to avoid it even after being instructed not to (Bromme et al., 2012). However, it is also possible that novice tutors lack competence in giving explicit feedback (Chi et al., 2004). Further research should examine the connection between politeness- and tutoring-related communication acts in order to establish the relative impact of tutor competence and politeness considerations on the use of different instructional strategies.

The differences in communication styles caused by the manipulation resulted in only minor changes in tutees' appraisals of the tutoring situation. The tutees did not assess their tutors' facework differently. Those in the no-politeness condition arguably indi-

cated feeling closer to their tutors than those in the politeness condition. This runs counter to the predictions of politeness theory: Brown and Levinson (1987) defined negative politeness strategies as *social brakes* that aim at maintaining a distance to the recipient and thereby preserving their autonomy. Tutees instructed by a “nonpolite” tutor also felt that they had learned more from the tutoring and would more readily work with their tutor again. This is especially interesting, because the actual learning outcomes did not differ. One explanation for these findings could be the content domain: Statistics is an aversive topic for many students (Baloğlu, 2003), and the tutees probably felt direct instruction to be appropriate in this well-structured domain (Buehl & Alexander, 2005). Moreover, tutors instructed to disregard politeness seemed to be less shy about sharing their knowledge, giving many more hints and correct solutions. It is conceivable that this elicited a feeling of knowing on the side of the tutees.

The finding that the different politeness instructions did not lead to differences in learning outcomes cannot be explained by a ceiling effect, because there was sufficient variance in the quality of tutees’ solutions of the ANOVA transfer task in both conditions. Therefore, our results fail to refute the notion that politeness does not hinder learning and can be beneficial (Kerssen-Griep et al., 2003). It is likely that learners “look behind” polite redresses, and that politeness—as long as it is not evasive—can help establish positive relations between tutor and tutee.

Limitations and Future Research

Whereas the experimental manipulation succeeded in changing the proportion of some politeness strategies and instructional behaviors in tutorial communication, it did not yield all the expected effects—especially on the tutees’ side. Multiple correlational and experimental studies have found that learners perceive polite instructors more positively on instructionally relevant variables (Jucks et al., 2012; Trees, Kerssen-Griep, & Hess, 2009; Witt & Kerssen-Griep, 2011). The fact that the current study did not show these effects supports the assumption that the manipulation was ineffective in bringing about significant changes in politeness that were also sufficiently salient for tutees. It is conceivable that the single, computer-mediated interactions in this study were too short for the tutees to form a detailed impression of their tutor. It would be interesting to observe longer interactions and gain a better evaluation of the social consequences of differentially polite communication.

If tutees are usually aware of politeness in tutoring, it is likely that tutors are as well. However, our study does not give insight on whether tutors avoid necessary instructional moves because they find them too imposing. We are not aware of any study that has investigated how the appropriateness of face-threatening moves in tutorial communication is assessed from a tutor’s perspective. Further research should endeavor to elucidate whether tutors acknowledge “the dialectical tension between effectiveness and appropriateness” (Carson & Cupach, 2000, p. 229): Do they perceive instructional FTAs as sanctioned behavior in tutoring interactions and hence utter them explicitly (cf. Hirokawa, Mickey, & Miura, 1991), or do they unproductively avoid them as has been observed in other contexts (Bonnefon et al., 2011)? Answering this question will shed light on whether better politeness instructions could bring about more effective tutoring. This is also relevant for other

instructional contexts such as expert–layperson communication: If doctors do not construe their role as teaching when explaining a diagnosis to their patients, evasive politeness can become a problem.

A limitation of our study is that although it required tutors and tutees to interact over a longer period, we analyzed politeness only on the level of tutors’ utterances. Because it is most likely that politeness is also negotiated discursively (Conlan, 2005; Holtgraves, 2005), a procedural perspective on data like ours could possibly deliver further insight into the processes governing tutorial interaction on this level. Some tutoring research has explored more complex analyses that could accomplish this (D’Mello, Olney, & Person, 2010).

Nonetheless, despite these shortcomings, we consider that the present study has delivered evidence for the important role of facework in instructional communication. Tutorial communication is in most cases inherently face-threatening. Using different measures, contexts, and methods to study the social, cognitive, and motivational influences on using politeness as well as the effects of using politeness on both sides—tutors and tutees—will help determine which kinds of effects are to be expected from which kind of tutor communication, and make it possible to fine-tune the kinds of instructions given to tutors.

References

- Baloğlu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences*, 34(5), 855–865. doi:10.1016/S0191-8869(02)00076-4
- Bonnefon, J.-F., Feeney, A., & De Neys, W. (2011). The risk of polite misunderstandings. *Current Directions in Psychological Science*, 20(5), 321–324. doi:10.1177/0963721411418472
- Bromme, R., Brummernhenrich, B., Becker, B.-M., & Jucks, R. (2012). The effects of politeness-related instruction on medical tutoring. *Communication Education*, 61(4), 358–379. doi:10.1080/03634523.2012.691979
- Bromme, R., Jucks, R., & Runde, A. (2005). Barriers and biases in computer-mediated expert–layperson-communication. An overview and insights into the field of medical advice. In R. Bromme, F. W. Hesse, & H. Spada (Eds.), *Barriers, biases and opportunities of communication and cooperation with computers—and how they may be overcome* (pp. 89–118). New York, NY: Springer.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge, England: Cambridge University Press.
- Buehl, M. M., & Alexander, P. A. (2005). Motivation and performance differences in students’ domain-specific epistemological belief profiles. *American Educational Research Journal*, 42(4), 697–726. doi:10.3102/00028312042004697
- Carson, C. L., & Cupach, W. R. (2000). Facing corrections in the workplace: The influence of perceived face threat on the consequences of managerial reproaches. *Journal of Applied Communication Research*, 28(3), 215–234. doi:10.1080/00909880009365572
- Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, 6(3), 271–315. doi:10.1207/s15327809jls0603_1
- Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students’ understanding accurately? *Cognition and Instruction*, 22(3), 363–387. doi:10.1207/s1532690xci2203_4
- Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19(2), 237–248. doi:10.3102/00028312019002237
- Conlan, C. J. (2005). Face threatening acts, primary face threatening acts,

- and the management of discourse. In R. T. Lakoff & S. Ide (Eds.), *Broadening the horizon of linguistic politeness* (pp. 65–83). Amsterdam, the Netherlands: Benjamins.
- Cromley, J. G., & Azevedo, R. (2005). What do reading tutors do? A naturalistic study of more and less experienced tutors in reading. *Discourse Processes*, 40(2), 83–113. doi:10.1207/s15326950dp4002_1
- Cupach, W. R., & Carson, C. L. (2002). Characteristics and consequences of interpersonal complaints associated with perceived face threat. *Journal of Social and Personal Relationships*, 19(4), 443–462. doi:10.1177/0265407502019004047
- D'Mello, S., Lehman, B., & Person, N. (2010). Expert tutors' feedback is immediate, direct, and discriminating. In H. W. Guesgen & R. C. Murray (Eds.), *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference* (pp. 595–604). Menlo Park, CA: AAAI Press.
- D'Mello, S., Olney, A., & Person, N. (2010). Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining*, 2(1), 1–37.
- Echterhoff, G., Lang, S., Krämer, N., & Higgins, E. T. (2009). Audience-tuning effects on memory: The role of audience status in sharing reality. *Social Psychology*, 40(3), 150–163. doi:10.1027/1864-9335.40.3.150
- Erceg-Hurn, D. M., & Miroseovich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591–601. doi:10.1037/0003-066X.63.7.591
- Goffman, E. (1967). *Interaction ritual: Essays on face-to-face interaction*. Oxford, England: Aldine.
- Graesser, A. C., D'Mello, S., & Cade, W. (2011). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 408–426). New York, NY: Routledge.
- Graesser, A. C., Person, N., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9(6), 495–522. doi:10.1002/acp.2350090604
- Guay, F., Vallerand, R. J., & Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion*, 24(3), 175–213. doi:10.1023/A:1005614228250
- Gueorguieva, R., Rosenheck, R., & Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis*, 52(12), 5344–5355. doi:10.1016/j.csda.2008.05.030
- Hijazi, R. H., & Jernigan, R. W. (2009). Modeling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics*, 4(1), 77–91.
- Hirokawa, R. Y., Mickey, J., & Miura, S. (1991). Effects of request legitimacy on the compliance-gaining tactics of male and female managers. *Communication Monographs*, 58(4), 421–436. doi:10.1080/03637759109376239
- Holtgraves, T. (2005). Social psychology, cognitive psychology, and linguistic politeness. *Journal of Politeness Research*, 1(1), 73–93. doi:10.1515/jplr.2005.1.1.73
- Howitt, D., & Cramer, D. (2008). *Introduction to SPSS in psychology: For Version 16 and earlier*. Harlow, England: Pearson Education.
- Jessmer, S. L., & Anderson, D. (2001). The effect of politeness and grammar on user perceptions of electronic mail. *North American Journal of Psychology*, 3(2), 331–346.
- Jucks, R., Brummehenrich, B., & Päuler, L. (2012). "I need to be explicit: You're wrong": Impact of face threats on social evaluations in online instructional communication. Manuscript submitted for publication.
- Kallia, A. (2005). Directness as a source of misunderstanding: The case of requests and suggestions. In R. T. Lakoff & S. Ide (Eds.), *Broadening the horizon of linguistic politeness* (pp. 217–234). Amsterdam, the Netherlands: Benjamins.
- Kerssen-Griep, J., Hess, J. A., & Trees, A. R. (2003). Sustaining the desire to learn: Dimensions of perceived instructional facework related to student involvement and motivation to learn. *Western Journal of Communication*, 67(4), 357–381. doi:10.1080/10570310309374779
- Kerssen-Griep, J., Trees, A. R., & Hess, J. A. (2008). Attentive facework during instructional feedback: Key to perceiving mentorship and an optimal learning environment. *Communication Education*, 57(3), 312–332. doi:10.1080/03634520802027347
- Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135–158). San Diego, CA: Academic Press. doi:10.1016/B978-012064455-1/50010-5
- Person, N., Kreuz, R. J., Zwaan, R. A., & Graesser, A. C. (1995). Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and Instruction*, 13(2), 161–188. doi:10.1207/s1532690xcil302_1
- Pighin, S., & Bonnefon, J.-F. (2011). Facework and uncertain reasoning in health communication. *Patient Education and Counseling*, 85(2), 169–172. doi:10.1016/j.pec.2010.09.005
- Trees, A. R., Kerssen-Griep, J., & Hess, J. A. (2009). Earning influence by communicating respect: Facework's contributions to effective instructional feedback. *Communication Education*, 58(3), 397–416. doi:10.1080/03634520802613419
- VanLehn, K., Siler, S. A., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3), 209–249. doi:10.1207/S1532690XCI2103_01
- Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, 66(2), 98–112. doi:10.1016/j.ijhcs.2007.09.003
- Witt, P., & Kerssen-Griep, J. (2011). Instructional feedback I: The interaction of facework and immediacy on students' perceptions of instructor credibility. *Communication Education*, 60(1), 75–94. doi:10.1080/03634523.2010.507820
- Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist*, 43(1), 49–64. doi:10.1080/00461520701756420

(Appendix follows)

Appendix

Instructions for Tutors (Translated From German)

These are the instructions for eliciting the experimental manipulation given to tutors before the tutoring sessions. The three

bulleted hints at the end of the instructions are the three phrases shown as prompts during the session.

Politeness condition	No-politeness condition
<p>Please note that learners will generally <i>feel offended if their autonomy is restricted and they are not free to make their own decisions.</i></p> <p>Therefore, when explaining these processes, it helps to <i>give learners the impression that their freedom of action is unrestricted.</i></p> <p>Try to <i>let them know that they can introduce their own ideas and make their own decisions on how to react to feedback.</i></p>	<p>Please note that learners will generally <i>feel offended if they do not understand what they are supposed to do or what their tutor means!</i></p> <p>Therefore, when explaining these processes, it helps to <i>communicate clearly and frankly.</i></p> <p>Try to <i>point out clearly what needs to be done next.</i></p>
<p>This is particularly important when you are phrasing requests or correcting the learner. Please pay attention to the following hints:</p> <ul style="list-style-type: none">• Phrase your <i>requests noncoercively!</i> Tutees should feel that they are free to choose whether to respond to your suggestions.• Phrase corrections in a way that <i>allows tutees to decide for themselves how they want to react to them!</i>• Phrase explanations in a way that <i>allows tutees to feel free to decide what they want to do with them!</i>	<p>Please pay attention to the following hints:</p> <ul style="list-style-type: none">• Phrase <i>requests clearly and explicitly!</i> Tutees should know exactly what they need to do.• Phrase corrections in a way that <i>lets tutees know what they have done wrong!</i>• Phrase explanations so that <i>tutees know exactly what you mean!</i>
<p>Communicating in this way ensures good understanding and makes it easier to solve the task correctly. These hints will appear on your screen from time to time during the session as reminders.</p>	

Received April 24, 2012

Revision received December 18, 2012

Accepted December 31, 2012 ■

Extraneous Perceptual Information Interferes With Children's Acquisition of Mathematical Knowledge

Jennifer A. Kaminski and Vladimir M. Sloutsky
The Ohio State University

Educational material often includes engaging perceptual information. However, this perceptual information is often extraneous and may compete with the deeper to-be-learned structure, consequently hindering either the learning of relevant structure or its transfer to new situations. This hypothesis was tested in 4 experiments in which 6- to 8-year-old children learned to read simple bar graphs. In some conditions, the bars were monochromatic (i.e., No Extraneous Information), whereas in other conditions, the bars consisted of columns of discrete countable objects (i.e., Extraneous Information). Results demonstrated that the presence of extraneous information substantially attenuated learning; participants tended to count the objects and failed to acquire the appropriate strategy. The interference effects decreased with age. These findings present evidence of how extraneous information affects learning of new mathematical knowledge. Broader implications of these findings for understanding the development of the ability to filter task-irrelevant information and for educational practice are also discussed.

Keywords: inhibition, learning, mathematics, number

Elementary educators are faced with a twofold challenge: They need to communicate content to students, and they need to keep students engaged in the process of learning. This dual responsibility may be particularly challenging when teaching mathematics, because mathematical concepts and procedures are often difficult for students to acquire (e.g., Brown & Burton, 1978; English & Halford, 1995).

One response to this challenge is to incorporate colorful familiar images into the learning material, with the goal of increasing children's engagement and linking the new mathematical content to some prior knowledge. For example, when the goal is for children to learn to read bar graphs, the bars might consist of columns of candies, animals, or other objects. Such colorful images are commonly encountered in elementary educational practice and can also be found in material intended for adults (e.g., "Chipotle store openings," 2010; "Favorite pizza toppings," n.d.). Furthermore, there is a common belief that such materials are helpful for learning. For example, we informally surveyed 16 kindergarten and elementary school teachers about material used to teach children to read bar graphs. In particular, we showed

teachers Graphs A and B in Figure 1 and asked them whether they would use similar graphs in their teaching and which of the two graphs would be more effective. All 16 teachers indicated that they would use graphs with columns of colorful objects in their teaching, with 14 of the teachers responding that such graphs would be more effective for teaching than graphs with monochromatic bars with no objects. Two teachers responded that they would not use the monochromatic graphs at all.

However, there is a note of caution: Although such added pictorial information may be visually appealing, such information may hinder, rather than facilitate, learning and/or transfer. This is because this information is often extraneous to the learning task, and it may also prompt well-learned strategies with unclear consequences on the to-be-learned ones. For example, items presented in Figures 1A and 1C may prompt counting, which is not an appropriate strategy for reading graphs. The inclusion of such extraneous information may be particularly problematic for children whose ability to control attention, filter irrelevant information, and inhibit prepotent responses is quite limited (see Hanania & Smith, 2010, for a review). Therefore, it is possible that such extraneous information may capture attention and prevent children from focusing on less salient to-be-learned structure or invite prepotent, well-learned strategies instead of newly presented strategies.

The goal of the present research was to examine how such extraneous perceptual information affects learning and how these effects change with development. We considered the case of learning to read bar graphs that had extraneous pictures similar to those in Figures 1A and 1C. Bar graphs present an interesting case for two reasons. First, bar graphs are a case of learning a relation (graphs depict a relation between two variables), and learning of relations is an important and challenging task during preschool and elementary school years (e.g., Goswami, 2001; Rattermann & Gentner, 1998). Second, bar graphs are an important real-life case

This article was published Online First December 17, 2012.

Jennifer A. Kaminski and Vladimir M. Sloutsky, Department of Psychology and Center for Cognitive Science, The Ohio State University.

This research was supported by Institute of Educational Sciences, U.S. Department of Education Grant R305B070407 and National Science Foundation Grant BCS-0720135. The opinions expressed herein are those of the authors and do not necessarily reflect the views of the Department of Education or the National Science Foundation. We thank Catherine Best and Anna Fisher for their helpful comments on this research.

Correspondence concerning this article should be addressed to Jennifer A. Kaminski, Department of Psychology, The Ohio State University, 225 Psychology Building, 1835 Neil Avenue, Columbus, Ohio 43210. E-mail: kaminski.16@osu.edu

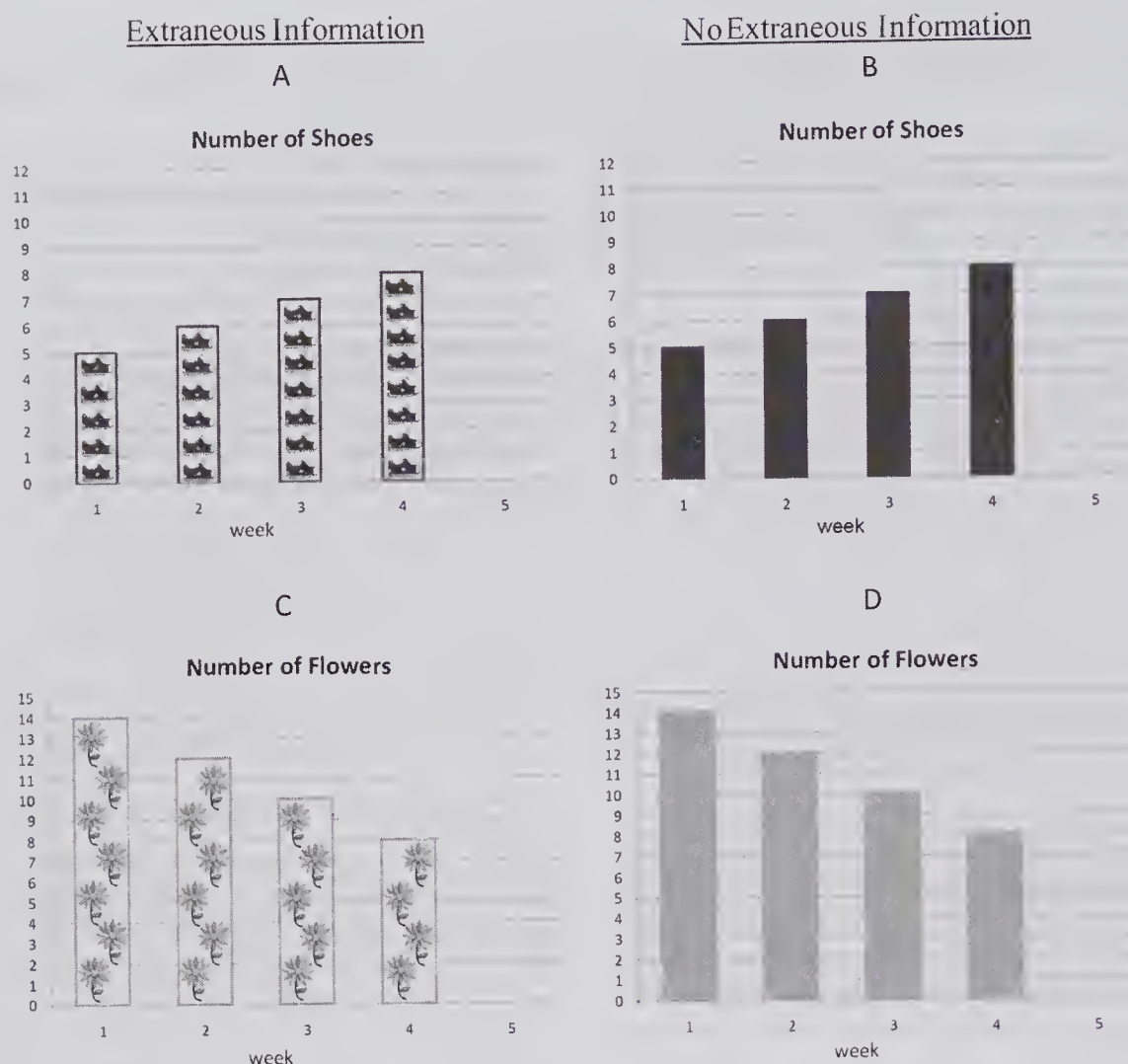


Figure 1. Sample stimuli: Graphs A and C were used in the extraneous information condition, and Graphs B and D were used in the no extraneous information condition. Graphs A and B were the example graphs shown in the Training phase, and Graphs C and D were test graphs shown in the Condition-Specific Test for each condition.

as they are part of elementary school curriculum (National Council of Teachers of Mathematics, 2000). Prior studies have investigated aspects of instruction that affect students' complex interpretations of graphs, such as noticing trends and interactions in the data (Kramarski, 2004; Shah & Hoeffner, 2002; Wainer, 1980, 1992). However, less is known about how variations in format affect young elementary students' basic graph reading ability, namely their ability to link a correct y-axis value with a specific x-axis value. In particular, how will children integrate and weigh task-relevant relational information and task-irrelevant (or extraneous) perceptual information?

Previous research suggests that children often have difficulty attending to relations and instead focus on superficial features (Gentner, 1988; Rattermann & Gentner, 1998). With development, they become increasingly capable of acquiring complex relational knowledge, such as basic arithmetic. Improvements in relational reasoning are largely driven by increases in domain knowledge (Goswami, 1992, 2001; Goswami & Brown, 1990) and in working memory capacity (Andrews & Halford, 2002; Halford, 1993; Richland, Morrison, & Holyoak, 2006). However, development per se does not guarantee successful relational reasoning. When relations are more complex than those involved in simple analo-

gies or metaphors (which is often the case with mathematical concepts), even adults can fail to recognize and transfer learned relations to novel contexts (Gick & Holyoak, 1980, 1983; Goswami, 1991; Kaminski, Sloutsky, & Heckler, 2008; Novick, 1988; Reed, Dempster, & Ettinger, 1985; Reed, Ernst, & Banerji, 1974).

Although complex relational knowledge is generally difficult to transfer, the format of the learning material can affect the likelihood of successful transfer. Undergraduate students who learned a novel mathematical concept from a generic, perceptually sparse learning format demonstrated better transfer than students who learned from a perceptually rich format (Sloutsky, Kaminski, & Heckler, 2005). There is also evidence that perceptually rich learning formats can hinder transfer of the nonmathematical relations (Goldstone & Sakamoto, 2003) as well as problem solving (McNeil, Uttal, Jarvin, & Sternberg, 2009). As we argued above, this is because examples used in this format communicate considerable extraneous information to the learner (see Kaminski & Sloutsky, 2011, for discussion), and this irrelevant information may be difficult for the learner to inhibit and filter.

The inclusion of extraneous information in learning of mathematics may be particularly detrimental for young children. First, mathematical knowledge is often relational in nature, and relations

are less salient than objects (e.g., Gentner, 1988). Adding extraneous superficial information increases the disparity in salience between the relevant relations and the superficial features. Second, filtering of irrelevant, potentially distracting information is particularly difficult for preschool and even elementary school children (Kemler, 1982; Shepp & Swartz, 1976; Smith & Kemler, 1978; see also Hanania & Smith, 2010). For example, Shepp and Swartz (1976) instructed 6- and 9-year-olds to sort items according to shape, with color being an irrelevant dimension. It was found that 6-year-olds (but not 9-year-olds) were slower when color varied independently of shape than when color covaried with shape or did not vary at all. Therefore, the task-irrelevant dimension affected performance of younger, but not older participants. Similarly, Napolitano and Sloutsky (2004) demonstrated that 4-year-olds have difficulty performing same-different discrimination with serially presented visual stimuli, when sounds accompanying these visual stimuli varied independently (see also Robinson & Sloutsky, 2004).

In addition, young children often experience difficulty inhibiting a previously learned response when a new response is needed (see Hanania & Smith, 2010; Plude, Enns, & Brodeur, 1994, for reviews). Although this ability improves with development (Davidson, Amso, Anderson, & Diamond, 2006), even adults are not immune to negative effects of such conflicts (Diamond & Kirkham, 2005; MacLeod, 1991). A variant of the difficulty to inhibit a well-learned response also transpires in the learning of mathematics. Children often apply well-learned, highly practiced procedures instead of to-be-learned ones. For example, children were found to apply familiar, yet inappropriate, arithmetic strategies to solve mathematical equivalence problems (McNeil, 2007; McNeil & Alibali, 2005) and whole number addition procedures to add fractions (Resnick & Ford, 1981).

In sum, in the context of extraneous information, successful learning (including learning of many mathematical concepts and procedures) may require attentional filtering and inhibitory control that may not be sufficiently developed in children. As a result, extraneous pictures that are intended to make learning material visually more appealing may actually hinder learning by either distracting from relational information or by prompting a well-learned, yet potentially inappropriate, strategy. Furthermore, given that the ability to filter irrelevant information and inhibit prepotent responses develops between preschool and late elementary school (Shepp & Swartz, 1976; see also Hanania & Smith, 2010; Plude et al., 1994, for reviews), we expect that these effects should reduce with age. Both hypotheses were tested in the four reported experiments.

Experiment 1

Method

Participants. Participants were 122 students recruited from public and private schools in suburbs of Columbus, Ohio, on the basis of returned parental consent forms. The majority of participants were Caucasian from middle-class families. Students were in kindergarten (20 girls, 20 boys; $M = 6.26$ years, $SD = 0.32$), first grade (18 girls, 24 boys; $M = 7.16$ years, $SD = 0.40$), and second grade (20 girls, 20 boys; $M = 8.24$ years, $SD = 0.38$).

Materials and design. Participants were randomly assigned to one of two between-subjects conditions (extraneous information or no extraneous information), which differed in the appearance of the graphs. The experiment consisted of three phases: Training, Condition-Specific Testing, and Novel Testing. Novel Testing was identical for both conditions. In all phases, participants were shown bar graphs representing quantities of different objects at different times, with quantities either increasing or decreasing with time. Time was always indicated on the x -axis, and quantity was always indicated on the y -axis.

Training. The Training phase consisted of the presentation of one example graph and one test graph. The appearance of the Training graphs differed across condition. In the no extraneous information condition, the bars were monochromatic (see Figure 1B), whereas in the extraneous information condition, the bars were filled with pictures of the objects whose quantities they represented (see Figure 1A). In the extraneous information condition, the number of objects in each of the bars was equal to the corresponding y -value (e.g., in Week 1, the number of shoes is five, and there are five shoes inside the corresponding bar). The experimenter demonstrated four separate readings of the Training example graph, and the participants were asked to make four readings (one for each bar) of the Training test graph.

Condition-Specific Testing. The Condition-Specific Testing phase followed the Training phase and consisted of three test graphs, for which participants made a total of 11 readings (four readings on each of the first and second graphs and three readings on the third graph). Critically, unlike in the Training phase, for the extraneous information condition, the number of objects in the bars did not equal the y -value; instead, the numbers of objects that appeared on each graph were proportionally related to the y -value as one third, one half, and one fourth of the y -values for the first, second, and third graphs, respectively (see Figure 1C). Therefore, responses clearly differentiated the children who correctly read the graph according to the y -axis from those who relied instead on the number of objects present. In the no extraneous information condition, the test graphs had monochromatic bars (see Figure 1D).

Novel Testing. The Novel Testing phase was condition-independent and consisted of two test graphs of a novel appearance. Participants in both conditions were presented with the same test graphs for which the bars were patterned with diagonal lines or small polka dots (see Figure 2). Participants were asked to make four separate readings of each test graph.

Procedure. Participants were tested individually at their schools by a female experimenter. All graphs were presented on a 15.6-in. laptop computer, and the experimenter recorded children's responses on paper. For each of the graphs, participants were read a situation involving quantities that were shown in the graph. In the Training phase, when presenting the example graph, the experimenter explicitly told the child that time was represented on the x -axis and stated each of the x -axis values. The experimenter also explicitly stated the y -axis values for this graph. For each of the four individual readings of the example graph, the experimenter pointed to the appropriate value on the x -axis, moved her finger upward over the bar of the graph, and then horizontally leftward to the y -axis to determine the corresponding y -value. The following is an excerpt from the script.

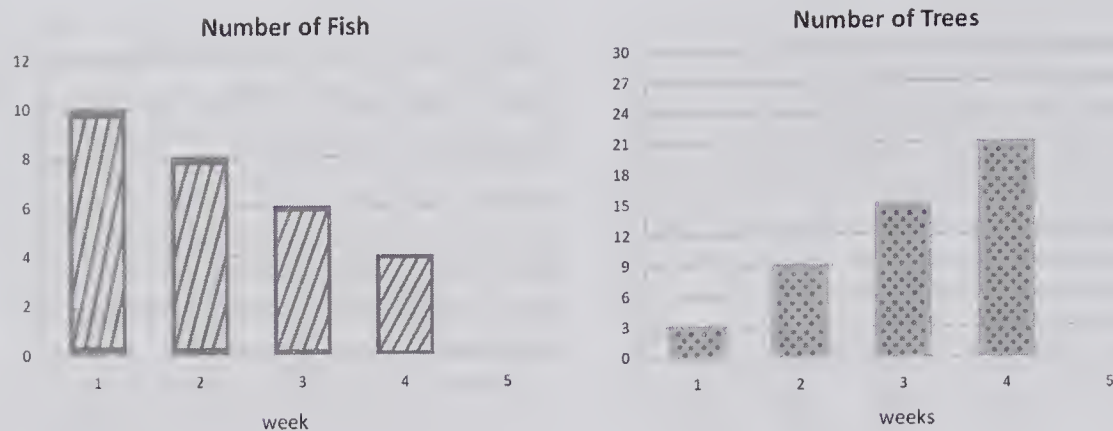


Figure 2. Novel Test Stimuli: Condition-independent Testing phase test graphs used in both the extraneous information and no extraneous information conditions.

These bottom numbers (the experimenter gestured across the bottom) stand for the week number. . . 1 for the 1st week, 2 for the 2nd week, 3rd week and 4th week (the experimenter pointed to each individually as she spoke). These numbers here (the experimenter gestured to the y-axis) stand for the number of shoes in the lost and found. So, for the 1st week (the experimenter pointed to the 1 on the x-axis) there were (the experimenter gestured upward over the bar then left to the y-axis) five shoes in the lost and found.

For each test graph in the Training phase, Condition-Specific Testing phase, and the Novel Testing phase, the experimenter read the scenario and explicitly told the participant what quantities the graph was showing (e.g., the number of flowers over several weeks). Then participants were asked to state the quantity for each indicated time point. No corrective feedback was given during training or testing.

Results

Responses within one unit of the correct y-value were considered to be correct, in order to credit participants who were using a correct graph reading strategy but made small errors in reading the value on the y-axis. Note that in this and other experiments reported here, the overall patterns remain the same and reported effects remain significant with or without this adjustment.

Training phase. In both the extraneous information and no-extraneous information conditions, participants in all age groups did well reading the Training phase test graph ($M = 100.0\%$, $SD = 0.0\%$ for second graders; $M = 95.0\%$, $SD = 22.4\%$ for first graders; and $M = 75.0\%$, $SD = 37.2\%$ for kindergarteners in the no extraneous information condition; and $M = 95.0\%$, $SD = 17.4\%$ for second graders; $M = 98.9\%$, $SD = 5.33\%$ for first graders; and $M = 86.3\%$, $SD = 25.0\%$ for kindergarteners in the extraneous information condition). An analysis of variance with condition and grade level as factors revealed a significant effect of grade level, $F(2, 116) = 7.96$, $p < .002$, with first and second graders being more accurate than kindergarteners (post hoc Tukey's, $ps < .003$). There was no significant effect of condition or interaction ($ps > .24$).

Note that for the Training phase test graph, the number of objects inside each bar in the extraneous information condition was equal to the y-value. Therefore, participants in the extraneous information condition could respond accurately either by counting

the objects in the appropriate bars or by reading the y-axis corresponding to the appropriate bars.

Condition-Specific Test phase. There were striking differences in accuracy between conditions on the Condition-Specific Test. For these graphs, the number of countable objects in the bars for the extraneous information condition was not equal to the y-value, resulting in incorrect responses from participants who counted these objects. Data across conditions and grade levels were not normally distributed. Therefore, we categorized participants on the basis of their predominant type of responses into one of three graph-reading strategies: correct, counting, or other. If greater than 50% of a participant's responses were correct, then he or she was categorized as a *correct* strategy user. The mean accuracy in this group for all ages and conditions was high, exceeding 89% (the same is true for the Novel Tests). Otherwise, if at least 50% of responses were based on the cardinality of the extraneous objects present in the bars, then the participant was categorized as a *counting* strategy user. Participants who did not fall into correct or counting categories were categorized as using *other* strategies. Responses from children in the *other* group appeared arbitrary. The mean accuracy for all ages and conditions in the latter two groups did not exceed 25% (the same is true for the Novel Tests).

Figure 3 presents the percentage of participants who used each of the strategy types and also the mean accuracy for all participants on the Condition-Specific Test. All first- and second-grade participants and 75% of kindergarten participants in the no extraneous information condition appropriately read the graphs. However, this was not true for the extraneous information condition in which 90% of kindergarteners and 72% of first graders responded by counting the objects. Differences in the number of correct strategy users were analyzed using an asymmetric log-linear analysis (Kennedy, 1992), with strategy type (correct or incorrect) as the dependent variable and condition and grade level as factors. Both condition and grade level were significant, $\chi^2(1, N = 122) > 55.3$, $p < .01$; and, $\chi^2(2, N = 122) > 26.4$, $p < .01$, respectively (.05 was added to cells with zero frequency). Older children were more accurate than younger children, $\chi^2(1, N = 82) > 3.09$, $p < .05$, one-tailed test between kindergarteners and first graders; and, $\chi^2(1, N = 82) > 5.56$, $p < .01$, one-tailed test between first graders and second graders.

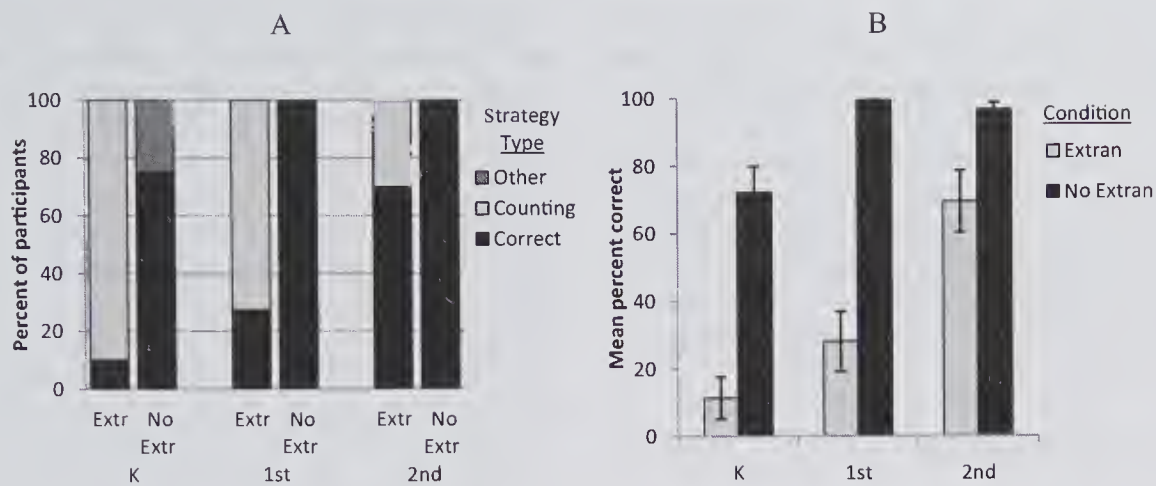


Figure 3. Performance on Condition-Specific Test in Experiment 1 split by condition and grade level. Panel A shows the percentage of participants performing each graph-reading strategy type. Panel B shows mean percent correct for all participants. Bars represent standard error of the mean. K = Kindergarten; 1st = first grade; 2nd = second grade; Extr = Extraneous; No Extr = No Extraneous; Extran = Extraneous; No Extran = No Extraneous.

Novel Test phase. Not only did many participants in the extraneous information condition count the discrete objects when present, but many of these participants also failed to appropriately read the Novel Test graphs (i.e., patterned bars). Figure 4 presents the percentage of participants who used each of the strategy types and also the mean accuracy for all participants on the Novel Test. To our surprise, many participants attempted to count the small stripes or polka dots on the bars, or they counted the horizontal lines present on the graph without considering the corresponding value on the y-axis. The remainder of participants who did not read the graphs accurately made arbitrary responses. However, in the no extraneous information condition, all first and second graders and 75% of kindergarteners accurately read these graphs. An asymmetric log-linear analysis with strategy type (correct or incorrect) as the dependent variable and condition and grade level as factors revealed differences across condition to be significant, $\chi^2(1, N = 122) > 11.1, p < .01$. Grade level was also a significant factor, $\chi^2(2, N = 122) > 25.6, p < .01$. First graders were more accurate

than kindergarteners, $\chi^2(1, N = 82) > 9.76, p < .01$, and second graders were marginally more accurate than first graders, $\chi^2(1, N = 82) = 2.67, p = .051$.

These results suggest that task-irrelevant, extraneous information interferes with learning, and these interference effects decrease with development. The presence of the discrete objects encouraged many kindergarten and first graders to count the objects instead of appropriately reading the graphs using the y-axis. Not only did these participants fail to use an appropriate strategy in the presence of these objects, many of them attempted to use counting strategies on the Novel Test graphs with the patterned bars. Others responded with arbitrary answers when there was no countable information present, suggesting that they failed to learn. Although it appears that the presence of these objects alone distracted participants from learning and using an appropriate graph-reading strategy, one could argue that the presence of the objects themselves is not harmful per se; rather, it is the conflict between the number of objects and the corresponding y-value that hindered

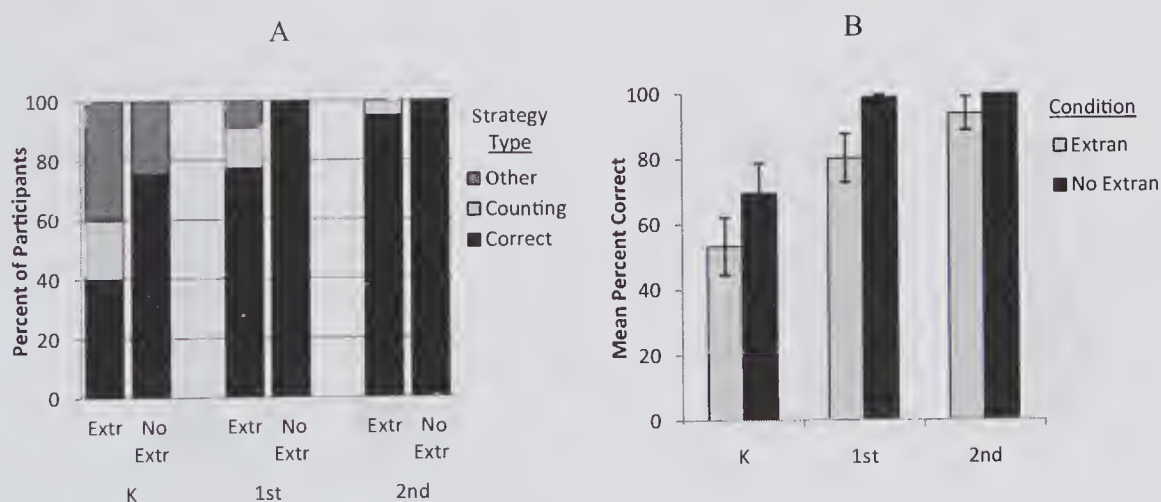


Figure 4. Performance on the Novel Test in Experiment 1 split by condition and grade level. Panel A shows the percentage of participants performing each graph-reading strategy type. Panel B shows mean percent correct for all participants. Bars represent standard error of the mean. K = Kindergarten; 1st = first grade; 2nd = second grade; Extr = Extraneous; No Extr = No Extraneous; Extran = Extraneous; No Extran = No Extraneous.

performance. Perhaps when these two pieces of information are in agreement, the objects may not hinder learning and may even facilitate learning by underscoring the fact that the bars represent quantities.

The purpose of Experiment 2 was to consider this possibility. In this experiment, the number of countable objects in a graph never conflicted with the corresponding y-value during the Training phase and the Condition-Specific Testing phase. Also, we only considered kindergarteners and first graders because the effect of the extraneous information was most pronounced for these two age groups.

Experiment 2

Method

Participants. Eighty-six students participated in the study. Students were in kindergarten (19 girls, 23 boys; $M = 6.14$ years, $SD = 0.32$) and first grade (26 girls, 18 boys; $M = 7.10$ years, $SD = 0.37$).

Materials and design. The material, design, and procedure were similar to those of Experiment 1, with one critical difference: Unlike Experiment 1, for the Condition-Specific Test in the extraneous information condition, the number of objects in each bar of the graphs equaled the corresponding y-value. Therefore, in the extraneous information condition, the cardinality of the extraneous objects never conflicted with the y-value. The actual numbers of the objects shown were the same as those in Experiment 1, but the y-values were set equal to these cardinalities. The same y-axis scale was used for both the extraneous information condition and the no extraneous information condition.

Results

Because the number of extraneous objects was equal to the y-value for the test graphs in the Training phase as well as the Condition-Specific Test phase, responses to these questions were not separated as in Experiment 1, but were analyzed together as the Condition-Specific Test responses.

Condition-Specific Test. In both the extraneous information and no extraneous information conditions, participants in both age groups did well reading the Condition-Specific graphs ($M = 98.8\%$, $SD = 2.63\%$ for first graders, and $M = 94.5\%$, $SD = 20.0\%$ for kindergarteners in the no extraneous information condition; and $M = 97.6\%$, $SD = 5.26\%$ for first graders, and $M = 92.7\%$, $SD = 8.35\%$ for kindergarteners in the extraneous information condition). An analysis of variance with condition and grade level as factors revealed no significant effect of condition and no significant interaction ($ps > .052$). There was a marginal effect of grade level, $F(1, 82) = 3.52$, $p = .064$, with first graders appearing slightly more accurate than kindergarteners. Therefore, similar to results with the Training Test graphs in Experiment 1, when the extraneous information does not conflict with the relevant y-values, participants accurately interpreted the graphs.

Novel Test. Although participants in both conditions were equally accurate reading the Condition-Specific Test graphs, the same was not true for reading the Novel Test graphs. The data were not normally distributed; therefore, participants were categorized on the basis of their predominant graph-reading strategy as correct, counting, or other (see Figure 5). The mean accuracy of correct strategy uses for both ages and conditions exceeded 95%. The mean accuracy for participants using the other two strategy types did not exceed 50%. In the no extraneous information condition, 100% of first graders and 91% of kindergarteners used a correct strategy to accurately read these graphs. However, in the extraneous information condition, only 77% of first graders and 45% of kindergarteners did so. The difference in the number of correct strategy users across condition was significant, asymmetric log-linear analysis, $\chi^2(1, N = 86) = 17.5$, $p < .01$. Additionally, first graders were more accurate than kindergarteners, $\chi^2(1, N = 86) = 6.62$, $p < .02$.

The results of Experiment 2 demonstrated that the hindering effects of the extraneous objects are not limited to situations when the cardinality information conflicts with the y-values. When cardinality is equal to the y-value, children were able to determine the correct values on the graphs. However, in the absence of such overt countable information (i.e., the patterned graphs), many

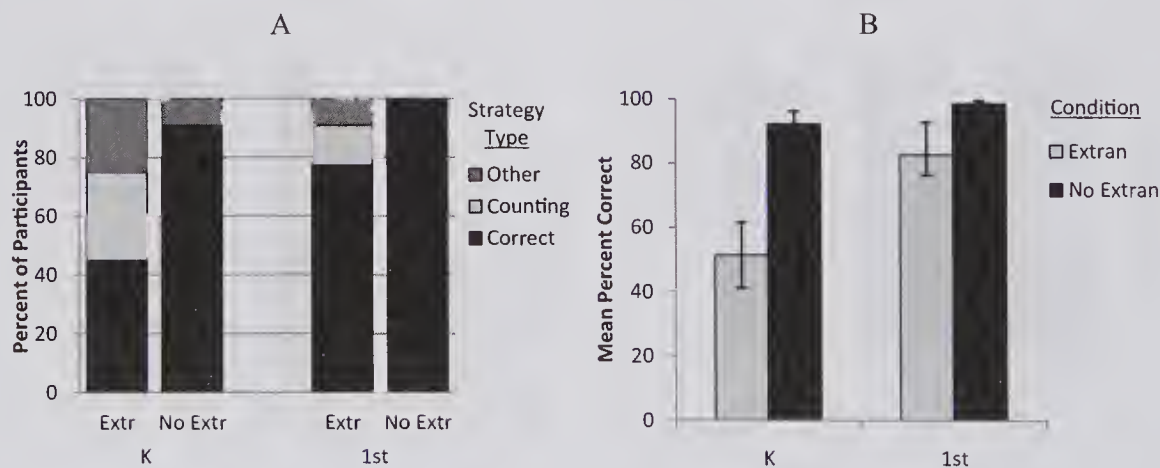


Figure 5. Performance on the Novel Test in Experiment 2 split by condition and grade level. Panel A shows the percentage of participants performing each graph-reading strategy type. Panel B shows mean percent correct for all participants. Bars represent standard error of the mean. K = Kindergarten; 1st = first grade; Extr = Extraneous; No Extr = No Extraneous; Extran = Extraneous; No Extran = No Extraneous.

children appear unable to correctly interpret the graphs. It seems that many children merely counted the objects they saw and failed to learn a correct graph-reading procedure.

The results of both Experiments 1 and 2 suggest that participants in the extraneous conditions did not learn to correctly read bar graphs. However, it may be that these participants actually did learn a correct graph-reading strategy, but noticed the equivalence of the y-values and the number of objects present on the example graph during the training phase and subsequently relied on a counting strategy in the presence of countable objects. They may also have overgeneralized this strategy to the novel test graphs, leading many to attempt to count aspects of the patterns (i.e., the strips or the dots). If this possibility is true, then the extraneous information may not have hindered learning; it may only have hindered performance by encouraging participants to count when countable objects or features were present. Furthermore, if participants in the extraneous information condition actually did learn how to read graphs correctly, then asking them to read graphs with monochromatic bars would be the optimal condition for them to do so. Graphs with monochromatic bars lack salient countable objects and features and therefore would eliminate or minimize the temptation to count.

The goal of Experiment 3 was to parse out the effect of the extraneous information on participants' performance and on participants' learning by testing all participants on graphs with monochromatic bars, patterned bars (i.e., stripes and dots as in the previous experiments), and bars with extraneous information (i.e., countable objects for which the number of objects conflicts with the y-value). If the extraneous information hinders only performance and not learning, then scores on the monochromatic graphs and countable object graphs should be comparable for participants in the extraneous information condition and participants in the no extraneous information condition. The absence of overtly countable extraneous objects on the monochromatic bars should encourage participants in either condition to use a proper graph-reading strategy if one had been learned. Alternatively, if the extraneous information hinders learning of the correct graph-reading strategy, scores on the monochromatic graphs should be lower for participants in the extraneous information condition than for those in the no extraneous information condition. Also, the presence of these objects on the graphs with extraneous information may encourage participants in the no extraneous information condition to use a counting strategy even though they have learned the appropriate graph-reading strategy. If this is the case, then it suggests that the extraneous information hinders performance even when a correct strategy was learned. Experiment 3 considered only kindergarteners as the effect of the extraneous information was most pronounced for this group of children.

Experiment 3

Method

Participants. Forty-four kindergarten students (21 girls, 23 boys; $M = 6.29$ years, $SD = 0.39$) participated in the study.

Materials and design. As in the previous experiments, there were two between-subject conditions, extraneous information and no extraneous information, which specified the appearance of the graphs that participants saw. In the extraneous information condi-

tion, the bars of the graphs had countable objects. In the no-extraneous information condition, the bars were monochromatic.

The experiment consisted of three phases: Training, Condition-Specific Testing, and General Testing. Training and Condition-Specific Testing were identical to Experiment 2. The General Testing was identical for the extraneous information condition and the no extraneous information condition and consisted of six test graphs with four readings each. Two of the six graphs presented patterned bars (i.e., the novel test graphs used in the previous experiments), two other graphs presented monochromatic bars, and the remaining two graphs presented bars filled with pictures of countable objects where the number of objects was not equal to the corresponding y-value. These six graphs were presented to participants in a random order.

Results

Responses within one unit of the correct y-value were considered to be correct, in order to credit participants who were using a correct graph-reading strategy but made small errors in reading the value on the y-axis.

Condition-Specific Test. In both the extraneous information and no extraneous information conditions, participants in both age groups did well reading the Condition-Specific graphs ($M = 92.7\%$, $SD = 12.7\%$ in the no extraneous information condition; and $M = 91.5\%$, $SD = 14.7\%$ in the extraneous information condition). There were no significant differences in accuracy between conditions (independent-samples t test), $t(42) = .293$, $p = .771$. As in the previous experiment, the extraneous information did not conflict with the y-values; therefore, participants in the extraneous information condition could arrive at the correct responses either through a correct graph-reading strategy or through counting the objects present.

General Test. The data for performance on the monochromatic graphs, patterned graphs, and countable object graphs was not normally distributed. Therefore, participants were categorized on the basis of their predominant graph-reading strategy as correct, counting, or other as done in the previous experiments. In both conditions and for all test graphs, the mean accuracy for participants categorized as correct strategy users exceeded 87%; the mean accuracy for participants in the other categories was below 50%.

Mean scores and frequency of strategies are presented in Figure 6. A repeated measures binary logistic regression was conducted to examine effects of training condition and test graph type on correct strategy use. There was a significant effect of condition, Wald $\chi^2(1, N = 44) = 12.5$, $p < .001$. For all three types of graphs, a higher percentage of participants in the no extraneous information condition than in the extraneous information condition used a correct graph-reading strategy, Pearson's $\chi^2(1, N = 44) > 5.93$, $ps < .02$. Test graph type also had a significant effect on strategy use, Wald $\chi^2(2, N = 44) = 15.8$, $p < .001$. Participants were more likely to use a correct strategy on the monochromatic than on the countable graphs, Wald $\chi^2(1, N = 44) = 15.3$, $p < .001$, and they were more likely to use a correct strategy on the patterned than on the countable graphs, Wald $\chi^2(1, N = 44) = 11.6$, $p < .002$.

When asked to read monochromatic graphs, many participants in the extraneous information condition failed to do so. Even in the absence of salient countable objects, 14% of participants re-

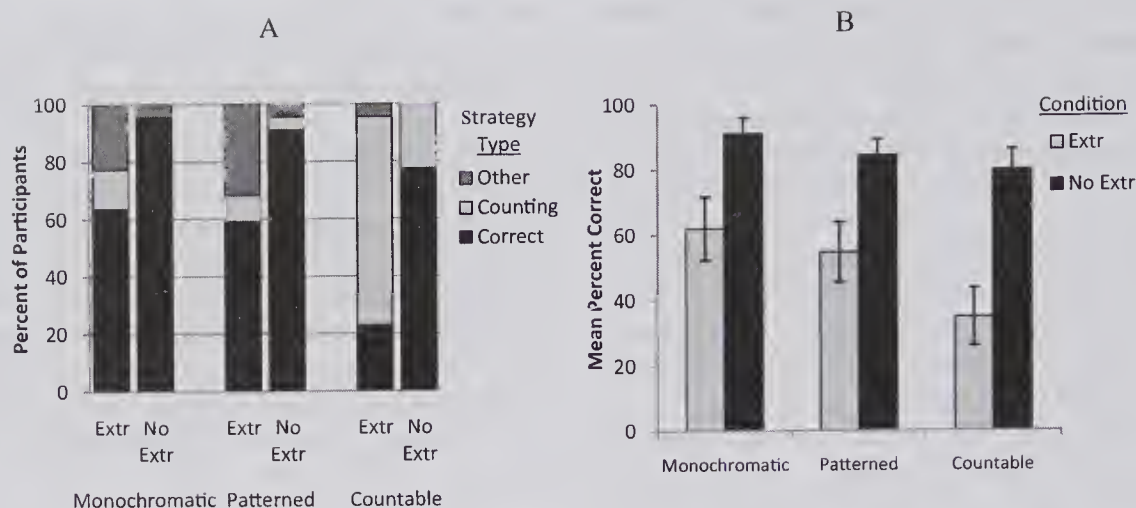


Figure 6. Performance on the General Test in Experiment 3 split by condition and graph type. Panel A shows the percentage of participants performing each graph-reading strategy type. Panel B shows mean percent correct for all participants. Bars represent standard error of the mean. Extr = Extraneous; No Extr = No Extraneous.

sponded by counting the horizontal lines through the graphs. Another 23% made arbitrary responses to these graphs. These results support the argument that the extraneous information hindered learning; it appears that 37% of participants in the extraneous information condition failed to learn a correct graph-reading strategy, whereas only 5% of participants in the no extraneous information condition did not learn. This difference was significant (Fisher's exact test, $p < .02$).

Although a majority of participants in the no extraneous information condition accurately read the monochromatic and patterned graphs, they were not all immune to the distraction of the countable objects; 22% responded to countable object graphs by counting the objects. The number of participants in the no extraneous information condition who counted was significantly fewer than the number in the extraneous information condition (77% counted), Pearson's $\chi^2(1, N = 44) = 11.0, p < .002$. These results suggest that although the presence of countable objects does encourage children to perform an incorrect strategy, they are less likely to do so if they initially learned a graph with monochromatic bars than if they initially learned a graph with extraneous objects on the bars.

The results of Experiment 3 suggest that the extraneous information hinders both learning and performance. On the monochromatic graphs, the percentage of participants in the extraneous information condition who accurately read the graphs was 32 points less than the percentage of participants in the no extraneous information condition (64% vs. 96%), suggesting that the extraneous information hindered learning. When comparing performance on monochromatic graphs with performance on countable object graphs for participants in the no extraneous information condition, there was a 19-point drop in percentage of participants responding correctly (96% vs. 77%), suggesting that the presence of the extraneous objects hindered performance, even for participants who acquired an accurate graph-reading strategy.

Although these results provide further evidence that the presence of the extraneous information distracted children from learning the relevant relational knowledge, there is an alternative explanation for the poor accuracy of participants in the extraneous information condition. It may be that these participants attended to

the relation between the x - and y -values on the example graph and noticed the equivalence of the y -values and the number of objects present. Because the number of objects always agreed with the experimenter's reported values, participants may have learned to rely on a counting strategy to read graphs and not on the appropriate relation between the x - and y -values. Therefore, it may be that the presence of the extraneous objects may not have hindered learning if participants did not observe that the number of objects corresponded with the correct readings on the example they were shown.

The purpose of Experiment 4 was to test this possibility. The design of this experiment was identical to that of Experiment 3 with one critical exception. The numbers of objects shown on the columns of the example graph did not equal the corresponding y -values. If performance was hindered not by the presence of the objects themselves but by the fact that the number of objects concurred with the correct readings on the example graphs, then performance in the extraneous information condition should be comparable to that of the no extraneous information condition.

Experiment 4

Method

Participants. Forty kindergarten students (20 girls, 20 boys; $M = 6.23$ years, $SD = 0.33$) participated in the study.

Materials and design. As in the previous experiments, there were two between-subject conditions, extraneous information and no extraneous information, which specified the appearance of the graphs that participants saw. The experiment consisted of three phases: Training, Condition-Specific Testing, and General Testing. Training and Condition-Specific Testing were similar to those of Experiments 2 and 3. However, the number of objects that appeared on the bars of the graphs in the Training and Condition-Specific Testing phases for the extraneous information condition was never equal to that of the corresponding y -values. The numbers of objects on the example graph were half of the corresponding y -values. The numbers of objects on the Condition-Specific Test graphs were one half, one third, one half, and one fourth of the

y-values for the first, second, third, and fourth graphs, respectively. The General Testing was the same for participants in both the extraneous information condition and the no extraneous information condition and identical to that of Experiment 3.

Results

Responses within one unit of the correct y-value were considered to be correct, in order to credit participants who were using a correct graph-reading strategy but made small errors in reading the value on the y-axis.

The data for performance on the condition-specific, monochromatic, patterned, and countable object graphs were not normally distributed. As in the previous experiments, participants were categorized on the basis of their predominant graph-reading strategy as correct, counting, or other for each test graph. In both conditions and for all test graphs, the mean accuracy for participants categorized as correct strategy users was high, exceeding 93%; the mean accuracy for participants in the other categories was below 50%. Figure 7 presents the frequencies of strategy types and the mean percent correct collapsed across all strategy types.

Unlike the previous experiment, there was a marked difference between the extraneous information and no extraneous information

conditions in accuracy on the Condition-Specific graphs. Significantly more participants in the no extraneous information condition used a correct strategy than those in the extraneous information condition (Fisher's exact test, $p < .03$).

For the monochromatic graphs, patterned graphs, and countable object graphs, a repeated measures binary logistic regression was conducted to examine the effects of training condition and test graph type on correct strategy use. The results reveal a significant effect of training condition, $\chi^2(1, N = 44) = 3.87, p < .05$, with no significant effect of test graph type, Wald $\chi^2(2, N = 44) = 3.31, p = .191$. The former effect indicated that across the test graph types, a higher number of participants in the no extraneous information condition than in the extraneous information condition used a correct strategy.

As in Experiment 3, the presence of the countable objects encouraged a small number of participants (10%) in the no-extraneous information condition to respond by counting the objects as opposed to correctly reading the graphs. However, 30% of participants in the extraneous information condition responded to these graphs by counting. The difference in frequency of correct strategy use between the two conditions approached significance (Fisher's exact test, $p = .06$).

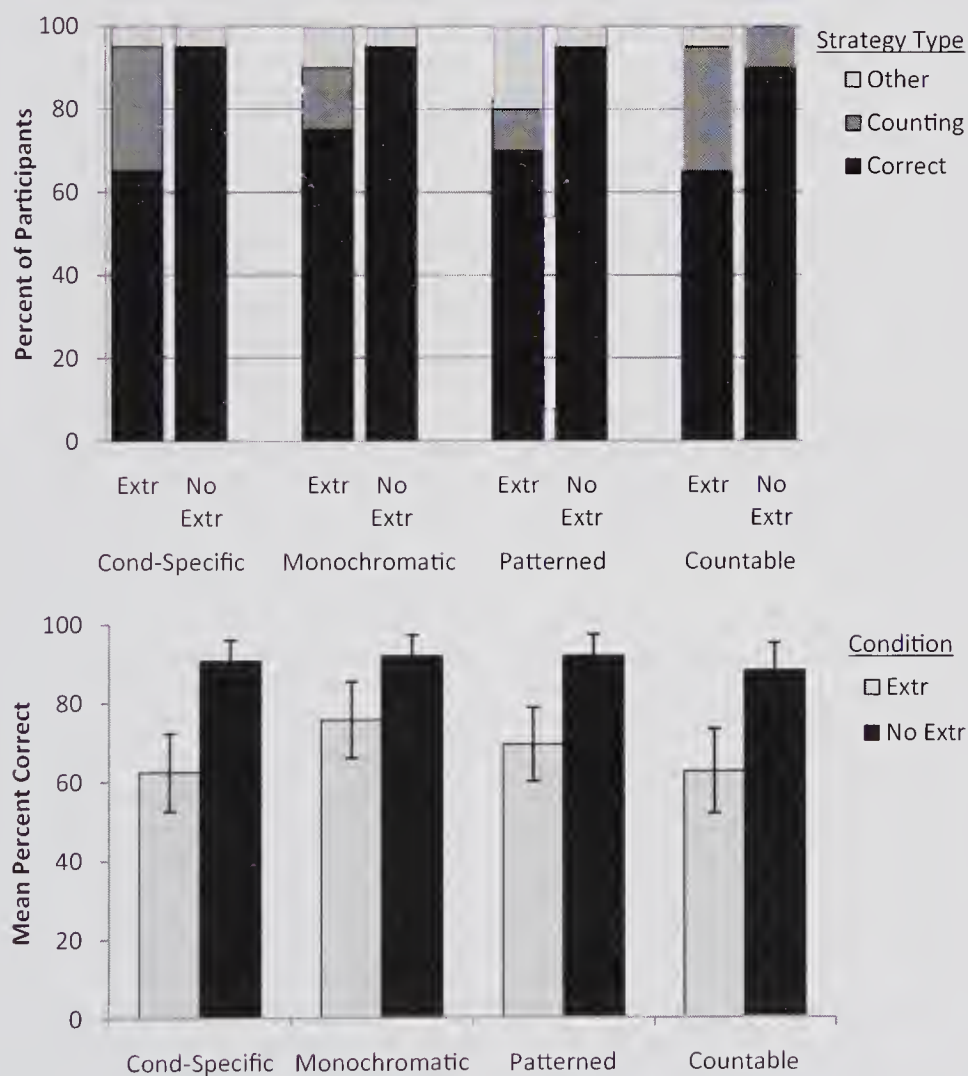


Figure 7. Performance on the Condition-Specific (Cond-Specific) Test and General Test in Experiment 4 split by training condition and test graph type. The top panel shows the percentage of participants performing each graph-reading strategy type. The bottom panel shows mean percent correct for all participants. Bars represent standard error of the mean. Extr = Extraneous; No Extr = No Extraneous.

Also similar to the previous experiment, many participants in the extraneous information condition failed to correctly read the monochromatic graphs; 15% responded by counting the horizontal lines through the graphs, and another 10% made arbitrary responses to these graphs. These results suggest that 25% of participants in the extraneous information condition failed to learn to correctly read bar graphs. In the no extraneous information condition, only 5% of participants did not read the graphs correctly. The difference in number of correct graph readers between the two conditions also approached significance (Fisher's exact test, $p = .09$).

Compared with the results of Experiment 3, more participants in the extraneous information condition of Experiment 4 learned to read the graphs correctly. To compare failure rates across the two experiments, the risk difference of failing to learn in the extraneous information condition versus failing to learn in the no extraneous information condition was used as a measure of effect size (see Ferguson, 2009). The risk difference in Experiment 3 was 32% (37% failure rate in the extraneous information condition vs. 5% in the no extraneous information condition), whereas the risk difference of Experiment 4 is 20% (25% failure rate in the extraneous information condition vs. 5% in the no extraneous information condition). The reduction in failure rate suggest that presenting graphs in which the number of objects did not equal the corresponding y -values enabled more children to learn than when the number of objects equaled the y -values.

General Discussion

The goal of the present research was to examine how extraneous perceptual information affects acquisition of simple mathematical knowledge by children. The results of Experiments 1 and 2 demonstrate that when discrete objects were depicted on the bars of a graph, the majority of children failed to read the graphs accurately and instead based their responses on the number of objects present. Not only did many participants respond by counting the objects when present, many of these children continued to use a counting strategy when shown the subsequent patterned graphs, basing their responses on extraneous features of the graph such as the number of stripes or dots. However, children who initially saw monochromatic bars accurately read these graphs.

The results also reveal a developmental trend with first and second graders (i.e., 7- and 8-year-olds) more accurately reading the graphs than kindergarteners (i.e., 6-year-olds), suggesting that older children are more capable of learning and using an appropriate procedure in the face of extraneous information. The better performance of the older participants is likely to stem from two factors. First, development leads to improvements in inhibitory control (Hanania & Smith, 2010; Plude et al., 1994; Shepp & Swartz, 1976), which is likely to increase children's ability to filter extraneous information and focus on relevant information. Second, the older children (i.e., first and second graders) may have had experience with bar graphs in school.

The results of Experiment 2 demonstrate that the difficulty introduced by the extraneous pictures is not limited to test situations when the extraneous and relevant information are in conflict. These findings suggest that it is not necessarily the conflict between information during testing that hindered children, but the conflict that arose during learning between possible procedures:

the to-be-learned graph-reading procedure and the well-learned counting procedure activated by the presence of the discrete objects.

Experiments 3 and 4 eliminate two alternative explanations for the poor performance of participants in the extraneous information conditions. Experiment 3 considered the possibility that the extraneous objects did not hinder learning, but hindered performance by encouraging a counting strategy when salient countable objects or features were present. Participants were tested on graphs with simple monochromatic bars as well as graphs with countable objects. The results were that performance in both the extraneous information and the no extraneous information conditions was better on graphs with monochromatic bars than on the graphs with countable objects; this suggests that the presence of the objects can hinder performance even for those in the no extraneous information condition who demonstrated correct graph reading. In addition, performance on the graphs with monochromatic bars was significantly lower in the extraneous information condition than in the no extraneous information condition, suggesting that the presence of the objects does hinder learning.

The goal of Experiment 4 was to examine the possibility that learning was not hindered by the presence of the objects per se but rather by the fact that the numbers of objects shown on the example graph always equaled the corresponding y -values. It may be that participants in the extraneous information condition did initially attend to the relation between the x - and y -values, but relied instead on a counting strategy because they observed that such a strategy led to correct readings of the example graph. The results demonstrate that when participants were shown an example graph for which they could not rely on counting the numbers of objects to arrive at a correct reading, subsequent performance was still lower in comparison to participants who were shown an example graph with monochromatic bars. These results provide further support for the argument that the extraneous information hindered learning.

Although graph-reading performance in Experiment 4 was lower for participants in the extraneous information condition than those in the no extraneous information condition, a comparison of risk differences (i.e., a measure of failure rate attributable to the extraneous information) between Experiments 3 and 4 suggests that the negative effect of the extraneous information is attenuated when this information is in conflict with the alternative strategy (i.e., counting). It may be that participants who noticed the conflict between the number of objects and the corresponding y -value redirected their attention to the relevant relation between the x -values and the y -values. These findings are perhaps counterintuitive because they suggest that in some situations, extraneous information that aligns with to-be-learned relational information may have a greater negative effect on learning than extraneous information that conflicts with to-be-learned relational information. Although creating a conflict between the relevant and extraneous information produced successful learning for many participants, this type of learning material does not appear to be widely used by teachers. In our informal survey of teachers who we mentioned earlier, only four of the 16 teachers indicated that they would use such graphs in their teaching.

Taken together, the results of all four experiments suggest that successful learning of mathematical procedures requires sufficient attention be allocated to the relevant underlying relations. Extra-

neous information included in the learning material can capture the learner's attention and divert it from these relations during learning. Not only can extraneous information hinder learning, it can also hinder subsequent performance. However, children who learned in the absence of extraneous information were more resistant to the hindering effects of extraneous information during testing than those who learned in the presence of extraneous information. In the present experiments, extraneous information came in the form of discrete, countable objects, which might intuitively seem to support learning because the number of these objects corresponded with the to-be-learned responses (in Experiments 1–3). However, it appears that the extraneous information activated a well-learned counting procedure that children used instead of learning a new procedure.

By elementary school, children are very familiar with the process of counting to determine cardinality. Understanding of the relationship between counting and cardinality emerges gradually over the preschool years (Baroody & Price, 1983; Briars & Siegler, 1984; Fuson, 1988; Gelman & Gallistel, 1978; Mix, 2002; Wynn, 1990, 1992). In school, instruction on arithmetic concepts often involves enumeration and counting of sets of discrete objects (see Van De Walle, 2007, for examples). Therefore, the process of counting to determine cardinality is very well practiced. This process can promote important mathematical understanding. At the same time, there may be instances, such as those of the present study, when counting and cardinality conflict with correct mathematical knowledge and procedures. This possibility has also been noted by other researchers who have suggested that well-learned counting knowledge may interfere with children's proportional reasoning (Boyer, Levine, & Huttenlocher, 2008). More generally, in the course of acquiring mathematical knowledge, children sometimes apply well-learned, highly practiced, but inappropriate procedures instead of to-be-learned ones (e.g., when solving equivalence problems and when performing fraction arithmetic; see McNeil, 2007; McNeil & Alibali, 2005; Resnick & Ford, 1981). However, unlike the situations with equivalence problems and fraction arithmetic, the temptation to apply a familiar but inappropriate strategy to bar graph reading can be easily avoided by removing the source of this temptation.

It could be argued that in the case of teaching children to read bar graphs, the inclusion of columns of objects is intended to scaffold learning with students initially instructed to read pictograms for which the number of objects present intentionally represents the relevant quantities. The present study suggests two points. First, if such objects are included in instructional material, explicit measures need to be taken to direct children's attention to the relevant relation between the x and y variables. Such measures may come in the form of creating a conflict between possible responses, as in Experiment 4, or possibly in the form of phasing the extraneous information out the learning material. Progressive fading of perceptual details has been shown to facilitate transfer of knowledge (Goldstone & Son, 2005). Second, the high level of accuracy in the no extraneous information conditions suggests that such a scaffold is not necessary for successful learning.

The present study revealed that extraneous perceptual information hinders learning of reading graphs, especially in kindergarten (i.e., 6- to 7-year-old) children. However, several

related issues remain unknown. First, in the present study we examined children's most basic understanding of bar graphs, namely, how to associate a particular independent variable with the appropriate dependent variable. The effects of extraneous information on other aspects of bar graph reading, such as noticing trends in the data or interpreting contextual significance, would need further experimental consideration. In addition, the extent to which the extraneous information may affect participants' generalizability of the learned graph-reading procedure is not known. For example, can participants generalize the learned procedure to appreciably different bar graphs, such as graphs with horizontal bars? Second, the extraneous information examined in this study was both countable and perceptually rich. To know the extent to which the demonstrated effect is attributed to each of these factors independently would need further research. Clearly, the effect is attributed in part to the presence of discrete objects; otherwise, there would be nothing to count. We also expect that the perceptual richness of the objects also contributes to the effect by making the objects very salient, as opposed to less salient countable circles for example.

The present results have important implications for the design of educational material. These findings underscore the importance of considering children's limited attentional capacities when designing and introducing learning material. Those who design material need to consider the possibility that inclusion of extraneous perceptual information may divert attention from the to-be-learned information. The consequence of children not acquiring the new knowledge may not be immediately recognizable by teachers in situations when the extraneous information is not in conflict with to-be-learned procedures (e.g., in the present study, when cardinality equaled the y -value). Instead, failure in learning can appear later when children are presented with material from which they cannot rely on another procedure. Therefore, it is important that the designers of instructional material, including textbooks and lesson plans, not simply rely on intuition as to what features may seem desirable or visually pleasing. They should recognize a priori the potential pitfalls of including such extraneous information in learning material intended for children whose ability to inhibit extraneous information is still developing.

References

- Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, 45, 153–219. doi:10.1016/S0010-0285(02)00002-6
- Baroody, A. J., & Price, J. (1983). The development of the number-word sequence in the counting of three-year-olds. *Journal for Research in Mathematics Education*, 14, 361–368. doi:10.2307/748681
- Boyer, T., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology*, 44, 1478–1490. doi:10.1037/a0013110
- Briars, D., & Siegler, R. S. (1984). A featural analysis of preschoolers' counting knowledge. *Developmental Psychology*, 20, 607–618. doi:10.1037/0012-1649.20.4.607
- Brown, J. S., & Burton, R. B. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, 2, 155–192. doi:10.1207/s15516709cog0202_4
- Chipotle store openings. (2010). In *Delta Sky Magazine*, p. 28. Retrieved from <http://msp.imirus.com/Mpowered/book/vds10/i12/p0>

- Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44, 2037–2078. doi:10.1016/j.neuropsychologia.2006.02.006
- Diamond, A., & Kirkham, N. (2005). Not quite as grown-up as we like to think: Parallels between cognition in childhood and adulthood. *Psychological Science*, 16, 291–297. doi:10.1111/j.0956-7976.2005.01530.x
- English, L. D., & Halford, G. S. (1995). *Mathematics education: Models and processes*. Hillsdale, NJ: Erlbaum.
- Favorite pizza toppings. (n.d.). In *National Council of Teachers of Mathematics Illuminations: Resources for Teaching Math*. Retrieved from <http://illuminations.nctm.org/LessonDetail.aspx?ID=L222>
- Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40, 532–538. doi:10.1037/a0015808
- Fuson, K. C. (1988). *Children's counting and concepts of number*. New York, NY: Springer-Verlag. doi:10.1007/978-1-4612-3754-9
- Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 59, 47–59. doi:10.2307/1130388
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306–355. doi:10.1016/0010-0285(80)90013-4
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38. doi:10.1016/0010-0285(83)90002-6
- Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology*, 46, 414–466. doi:10.1016/S0010-0285(02)00519-4
- Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealize simulations. *The Journal of the Learning Sciences*, 14, 69–110. doi:10.1207/s15327809jls1401_4
- Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development*, 62, 1–22. doi:10.2307/1130701
- Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Erlbaum.
- Goswami, U. (2001). Analogical reasoning in children. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 437–470). Cambridge, MA: MIT Press.
- Goswami, U., & Brown, A. L. (1990). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, 35, 69–95. doi:10.1016/0010-0277(90)90037-K
- Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.
- Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching. *Developmental Science*, 13, 622–635. doi:10.1111/j.1467-7687.2009.00921.x
- Kaminski, J. A., & Sloutsky, V. M. (2011). Representation and transfer of abstract mathematical concepts in adolescence and young adulthood. In V. Reyna (Ed.), *The adolescent brain: Learning, reasoning, and decision making* (pp. 67–93). Washington, DC: American Psychological Association. doi:10.1037/13493-003
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science*, 320, 454–455. doi:10.1126/science.1154659
- Kemler, D. G. (1982). Cognitive development in the school years: Foundations and directions. In J. Worrell (Ed.), *Psychological development in the elementary school years* (pp. 233–268). New York, NY: Academic Press.
- Kennedy, J. J. (1992). *Analyzing qualitative data: Log-linear analysis for behavioral research*. New York, NY: Praeger.
- Kramarski, B. (2004). Making sense of graphs: Does metacognitive instruction make a difference on students' mathematical conceptions and alternative conceptions? *Learning and Instruction*, 14, 593–619. doi:10.1016/j.learninstruc.2004.09.003
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163–203. doi:10.1037/0033-2909.109.2.163
- McNeil, N. M. (2007). U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Developmental Psychology*, 43, 687–695. doi:10.1037/0012-1649.43.3.687
- McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development*, 76, 883–899. doi:10.1111/j.1467-8624.2005.00884.x
- McNeil, N. M., Uttal, D. H., Jarvin, L., & Sternberg, R. J. (2009). Should you show me the money? Concrete objects both hurt and help performance on mathematics problems. *Learning and Instruction*, 19, 171–184. doi:10.1016/j.learninstruc.2008.03.005
- Mix, K. S. (2002). The construction of number concepts. *Cognitive Development*, 17, 1345–1363.
- Napolitano, A. C., & Sloutsky, V. M. (2004). Is a picture worth a thousand words? The flexible nature of modality dominance in young children. *Child Development*, 75, 1850–1870. doi:10.1111/j.1467-8624.2004.00821.x
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 510–520. doi:10.1037/0278-7393.14.3.510
- Plude, D. J., Enns, J. T., & Brodeur, D. (1994). The development of selective attention: A life-span overview. *Acta Psychologica*, 86, 227–272. doi:10.1016/0001-6918(94)90004-3
- Rattermann, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development*, 13, 453–478. doi:10.1016/S0885-2014(98)90003-X
- Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 106–125. doi:10.1037/0278-7393.11.1.106
- Reed, S. K., Ernst, G. W., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. *Cognitive Psychology*, 6, 436–450. doi:10.1016/0010-0285(74)90020-6
- Resnick, L. B., & Ford, W. W. (1981). *The psychology of mathematics for instruction*. Hillsdale, NJ: Erlbaum.
- Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, 94, 249–273. doi:10.1016/j.jecp.2006.02.002
- Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development*, 75, 1387–1401. doi:10.1111/j.1467-8624.2004.00747.x
- Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14, 47–69. doi:10.1023/A:1013180410169
- Shepp, B. E., & Swartz, K. B. (1976). Selective attention and the processing of integral and nonintegral dimensions: A developmental study. *Journal of Experimental Child Psychology*, 22, 73–85. doi:10.1016/0022-0965(76)90091-6
- Sloutsky, V. M., Kaminski, J. A., & Heckler, A. F. (2005). The advantage of simple symbols for learning and transfer. *Psychonomic Bulletin & Review*, 12, 508–513. doi:10.3758/BF03193796

- Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, 10, 502–532. doi: 10.1016/0010-0285(78)90009-9
- Van De Walle, J. A. (2007). *Elementary and middle school mathematics: Teaching developmentally*. Boston, MA: Pearson Education.
- Wainer, H. (1980). A test of graphicacy in children. *Applied Psychological Measurement*, 4, 331–340. doi:10.1177/014662168000400305
- Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher*, 21, 14–23.
- Wynn, K. (1990). Children's understanding of counting. *Cognition*, 36, 155–193.
- Wynn, K. (1992). Children's acquisition of number words and the counting system. *Cognitive Psychology*, 24, 220–251.

Received July 3, 2012

Revision received October 25, 2012

Accepted November 1, 2012 ■

Complex Problem Solving in Educational Contexts—Something Beyond *g*: Concept, Assessment, Measurement Invariance, and Construct Validity

Samuel Greiff and Sascha Wüstenberg
University of Luxembourg

Gyöngyvér Molnár
University of Szeged

Andreas Fischer and Joachim Funke
University of Heidelberg

Benő Csapó
University of Szeged

Innovative assessments of cross-curricular competencies such as complex problem solving (CPS) have currently received considerable attention in large-scale educational studies. This study investigated the nature of CPS by applying a state-of-the-art approach to assess CPS in high school. We analyzed whether two processes derived from cognitive psychology, knowledge acquisition and knowledge application, could be measured equally well across grades and how these processes differed between grades. Further, relations between CPS, general mental ability (*g*), academic achievement, and parental education were explored. Hungarian high school students in Grades 5 to 11 ($N = 855$) completed MicroDYN, which is a computer-based CPS test, and the Culture Fair Test 20-R as a measure of *g*. Results based on structural equation models showed that empirical modeling of CPS was in line with theories from cognitive psychology such that the two dimensions identified above were found in all grades, and that there was some development of CPS in school, although the Grade 9 students deviated from the general pattern of development. Finally, path analysis showed that CPS was a relevant predictor of academic achievement over and above *g*. Overall, results of the current study provide support for an understanding of CPS as a cross-curricular skill that is accessible through computer-based assessment and that yields substantial relations to school performance. Thus, the increasing attention CPS has currently received on an international level seems warranted given its high relevance for educational psychologists.

Keywords: complex problem solving, general mental ability, intelligence, MicroDYN, education

Improving students' minds is considered a major challenge in education. One way to achieve this is by enhancing students' problem-solving skills (Mayer & Wittrock, 2006), which are captured in their ability to solve novel problems. The importance of problem solving for success in life is also reflected in the Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Co-operation and Development

(OECD), which is allegedly the most comprehensive and important international large-scale assessment in existence today (e.g., OECD, 2004, 2010). The PISA studies aim to evaluate educational systems worldwide by assessing 15-year-olds' competencies in the key subjects of reading, mathematics, and science and also to evaluate more complex cross-curricular skills such as complex problem solving (e.g., OECD, 2010).

Specifically, cross-curricular complex problem solving (CPS) was assessed in more than half a million students in over 70 countries (e.g., OECD, 2009) in the current PISA 2012 cycle.¹ As an example of a typical CPS task in PISA 2012, imagine that you just bought your first mobile phone ever, you have never worked with such a device, and now you want to send a text message. Essentially, there are two things you need to do: (a) press buttons in order to navigate through menus and to get feedback on your actions and (b) apply this knowledge to reach your goal, that is, to send a text message. These aspects of CPS are also reflected in Buchner's (1995) definition:

This article was published Online First February 18, 2013.

Samuel Greiff and Sascha Wüstenberg, EMACS Unit, University of Luxembourg, Luxembourg-Kirchberg, Luxembourg; Gyöngyvér Molnár, Institute of Education, University of Szeged, Szeged, Hungary; Andreas Fischer and Joachim Funke, Department of Psychology, University of Heidelberg, Heidelberg, Germany; Benő Csapó, Institute of Education, University of Szeged, Szeged, Hungary.

This research was funded by grants supported by the Fonds National de la Recherche Luxembourg (ATTRACT; ASSKI21) and by the German Federal Ministry of Education and Research (LSA004 and 01JG1062). We are grateful to the Technology Based Assessment group at DIPF (<http://tba.dipf.de>) for providing the authoring tool CBA Item Builder and technical support.

Correspondence concerning this article should be addressed to Samuel Greiff, EMACS Unit, University of Luxembourg, 6 rue Richard Coudenhove Kalergi, 1359 Luxembourg-Kirchberg, Luxembourg. E-mail: samuel.greiff@uni.lu

¹ In PISA, the term *interactive problem solving* (OECD, 2010) is used. Other labels referring to the same construct are *dynamic problem solving*, which focuses on the aspect of systems to change dynamically (e.g., Greiff, Wüstenberg, & Funke, 2012) and *complex problem solving* (Dörner, 1986, 1990), which emphasizes the aspect of the underlying system's complexity. In the present article, we use the term *complex problem solving* (CPS), which is the most established in research.

Complex Problem Solving is the successful interaction with task environments that are dynamic (i.e., change as a function of user's intervention and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process. (p. 14)

Funke (2010) and Raven (2000) concluded that CPS requires a series of complex cognitive operations such as planning and implementing actions, model building, or self-regulation. Enhancing these cognitive operations is the goal of any educational system, or, as Mayer and Wittrock (2006) put it: "One of educational psychology's greatest challenges [is to help] students become better problem solvers" (p. 299). However, CPS research that combines assessment and theory is rather scarce. The present study contributes to research on the nature and validity of CPS by applying a state-of-the-art approach to assess CPS in high school students.

Complex Problem Solving and *g*

Research on general mental ability, now often referred to as psychometric *g*, was also initially educationally motivated. That is, when Alfred Binet and Théodore Simon (1904) developed the first psychometric tests of *g*, their starting point was to objectively identify students with learning disabilities who were in need of specially tailored education. Ever since then, no other construct has been as extensively and continuously validated in educational contexts. Specifically, based on the assorted existing empirical evidence, Reeve and Hakel (2002) concluded that there is a common mechanism underlying human mental processes labeled psychometric *g*. Only a few researchers have recently challenged this view by questioning the importance of *g* or by introducing alternative concepts such as practical intelligence (e.g., Lievens & Chan, 2010), social intelligence (e.g., Kihlstrom & Cantor, 2011), or emotional intelligence (e.g., Goleman, 1995). That is, the overwhelming conceptual and empirical evidence has supported the educational importance of *g* concerning manifold external criteria. The most impressive accumulation of evidence was provided by Ree and Carretta (2002), who related skills, personality, creativity, health, occupational status, and income to measures of *g*.

Theoretically, *g* is bolstered by the Cattell-Horn-Carroll (CHC) theory, which assumes that *g* is on a general level of cognitive ability (Stratum III), which in turn influences about 10 broad cognitive abilities on the second level (Stratum II). Narrow cognitive abilities are located on the lowest level (Stratum I; McGrew, 2009). CHC theory is considered particularly relevant to school psychologists and other practitioners for educational assessment and has received considerable attention in the educational arena. On a measurement level, strict requirements such as structural stability have been frequently shown to hold for tests of *g* (e.g., Taub & McGrew, 2004). Structural stability indicates that the construct does not change across groups and that test scores do not depend on the group to which the test is administered (Byrne & Stewart, 2006). This is a prerequisite for interpreting differences in mean performance (Cheung & Rensvold, 2002). In light of the overall empirical and theoretical evidence, it is not surprising that Reeve and Hakel (2002) consider *g* to be crucial in any educational context.

However, the predominant role of *g* in education has not been entirely undisputed. Whereas Sternberg (1984/2009) proposed a triarchic theory of intelligence composed of an analytical, a practical, and a creative component, Diaz and Heining-Boynton (1995) noted the relevance of alternative concepts such as CPS for students' education, thus, going beyond the idea that a single mental construct underlies cognitive performance and including more complex processes. The general rationale behind this idea is that despite the well-established predictive power of *g*, many questions about its nature remain unsolved (e.g., genetic endowment, environmental influence, different forms of intelligence; Neisser et al., 1996). In fact, *g*'s ability to predict nonacademic performance is considerable but far from perfect even after controlling for measurement error; thus, variance that may be accounted for by CPS is left unexplained (e.g., Rigas, Carling, & Brehmer, 2002). In this context, some studies on the relation between measures of CPS and *g* have yielded low relations. For example, Putz-Osterloh (1981) reported zero correlations between performance in the CPS scenario *Tailorshop* and the *Advanced Progressive Matrices* (Raven, 1962). Even though methodological issues might have caused this result, current findings have supported the distinction between *g* and CPS and have demonstrated the added value of CPS beyond *g* in different contexts (e.g., Danner, Hagemann, Schankin, Hager, & Funke, 2011; Wüstenberg, Greiff, & Funke, 2012).

Even during the early stages of CPS research, Dörner (1986) criticized the focus on the speed and accuracy of the capacity for basic information processing in measures of *g* (e.g., Raven, 2000) and suggested that a stronger emphasis be placed on the strategic and processing aspects of the mental processes involved in CPS. He proposed measuring complex cognitive processes in CPS to overcome the "out-of-touch-with-reality" issue that traditional intelligence tests suffer from (Dörner & Kreuzig, 1983). The broad conception of mental ability in CPS connects directly to the understanding of learning in the classroom. Mayer and Wittrock (2006) stated that a deep understanding of the nature of problem solving is needed if meaningful learning is to be fostered. Thus, going beyond current conceptualizations of *g*, meaningful learning and problem solving are closely related (Sternberg, 2000), and they are of great importance both to predict and to understand complex learning processes in classrooms (Mayer & Wittrock, 2006). Similar to Sternberg and his conception of intelligence (Sternberg, 1984/2009), the line of research on CPS that emerged around Dörner (1986) does not seriously object to the use of measures of *g* but suggests complementing them with additional measures such as CPS and its defining cognitive processes.

Complex Problem Solving in Cognitive Science

Mayer (2003) defined problem solving in general as transforming a given state into a goal state when no obvious method of solution is available. According to Funke and Frensch (2007), a problem solver has to overcome barriers by applying operators and tools to solve a problem. However, problem solving may take place in different educationally relevant domains, and a large body of research has been conducted in domain-specific areas such as mathematical, scientific, or technical problem solving (Sugrue, 1995). Besides these domain-specific approaches, the idea of domain-general processes generally involved in problem solving was taken up by the European line of research on complex problem

solving mentioned above (e.g., Dörner, 1986; Funke, 2001; Funke & Frensch, 2007).

This line of research assumes that domain-general processes are crucial when participants deal with an unknown and highly inter-related system (i.e., a complex problem) for the first time, although when dealing with the same problem repeatedly, domain-specific knowledge may be increasingly involved. That is, CPS research acknowledges that previous experiences or the problem context may influence CPS, but these aspects are not of elementary concern, and problems are designed to be solvable without domain-specific prior knowledge. In the tradition of Newell and Simon (1972), who described problem-solving behavior uncontaminated by domain-specific knowledge, CPS research aims to uncover general cognitive processes before a considerable amount of domain-specific prior knowledge is gathered and, thus, before problem solvers switch to more specialized strategies.

Generally, two main demands specify a problem solver's performance within the realm of CPS: knowledge acquisition and knowledge application (Funke, 2001). For instance, dealing with an entirely new mobile phone as outlined previously describes a specific situation that is typically considered to be a complex problem involving dynamic interaction with a yet-unknown system in order to (a) acquire knowledge and (b) use this knowledge for one's own purposes. Not only is this delineation into two main cognitive processes logical and widely applied when assessing CPS (e.g., Fischer, Greiff, & Funke, 2012; Funke, 2001; Kröner, Plass, & Leutner 2005), but it also connects to general research on (a) problem representation and (b) the generation of problem solutions.

Regarding problem representation, the Gestalt psychologist Duncker (1945) was the first to emphasize the importance of a sound problem representation, and Markman (1999) has further elaborated on this concept. According to Markman's elaboration, a representation begins with a description of the elements of a complex problem, the *represented world*, and a set of operators that can be used to relate these elements to each other, the *representing world*. Represented and representing worlds are usually predefined in CPS research, that is, the problems are well defined (represented world), and the set of operators available is limited and can be used only within given constraints (representing world; this setup is often found in educational contexts; Mayer & Wittrock, 2006). The elements of a complex problem (represented world) and the set of operators (representing world) are subsequently connected by a set of rules that are established while the problem solver attempts to penetrate the problem. This kind of task is often required of students in school and is at the core of the solver's task in CPS. It describes the process of building a problem representation. In the example above, a description of the problem (i.e., sending a text message) and the set of elements (i.e., inputs and outputs of the mobile phone) are predefined, but the connections between them are yet to be built. Finally, this needs to lead into a process that uses the representation that was established before the problem solution (Markman, 1999). It is this representational function that gives meaning to the representation (Novick & Bassok, 2005) and that constitutes the link between the problem representation (i.e., knowledge acquisition) and generating a problem solution (i.e., knowledge application).

Regarding the generation of a problem solution, algorithmic and heuristic strategies represent a common distinction between dif-

ferent types of solutions. Whereas algorithms are guaranteed to yield a solution, heuristics are usually applied when an exhaustive check of all possible moves is not efficient (Novick & Bassok, 2005). As this exhaustive check is scarcely possible in complex problems, it is safe to assume that the process of solving them is largely guided by heuristics such as a means-ends analysis (Newell & Simon, 1972). In fact, Greeno and Simon (1988) stated that problem solvers tend to prefer a means-ends analysis as the solution method when faced with novel problems that are relatively free of prior knowledge and in which well-defined goals are given. Often, when students face transfer problems in educational contexts, it is under exactly the condition that prior factual knowledge is of limited help in solving the problem at hand and that the available operators are clearly defined (Mayer & Wittrock, 2006).

Obviously, knowledge acquisition and knowledge application are closely entangled because a good representation is to a certain degree a necessary condition for establishing specific goals and for deducing interventions to solve a problem (Novick & Bassok, 2005). Thus, researchers in both of the two aforementioned fields have emphasized the importance of the respective aspect: Newell and Simon (1972) introduced the concept of a *problem space* in which the problem, its rules, and its states are represented, focusing on aspects of knowledge acquisition. By contrast, Markman (1999) considered the use of information essential and, thus, the process of knowledge application. Novick and Bassok (2005) stated that "although it is possible to focus one's research on one or the other of these components, a full understanding of problem solving requires an integration of the two" (p. 344). As it is widely acknowledged that representation and solutions interact with each other, the neglect of concrete efforts to converge these two lines of research has been surprising.

Measurement Approaches to Complex Problem Solving

A comprehensive assessment of the CPS dimension knowledge acquisition requires the active exploration of an unknown system, and assessment of knowledge application requires the immediate adaption to actions initiated by the system. Thus, by definition, the assessment of CPS is always computer-based, as the task changes interactively by itself or due to the user's intervention (Funke & Frensch, 2007), which cannot be assessed on a pencil-and-paper basis (Funke, 2001).

Consequently, computer-based microworlds (e.g., Gardner & Berry, 1995) were developed to reliably measure CPS performance. However, most efforts were overshadowed by severe measurement issues (cf. Greiff, Wüstenberg, & Funke, 2012; Kröner et al., 2005). It was only recently that multiple complex systems were introduced as another advance in the assessment of CPS (Greiff et al., 2012). In a multiple-complex-systems approach, time on each task is significantly reduced and tasks are directly scaled with regard to their difficulty (Greiff, 2012). Hence, in one testing session, problem solvers work on several independent tasks and are confronted with an entire battery of CPS tasks. In this manner, a wide range of tasks with varying difficulty can be employed, leading to increased reliability. Thus, the theoretically derived internal structure of CPS with its distinction between knowledge acquisition and knowledge application was able to be psychometrically confirmed for the first time with the

a name for a specific cat food) or without deep semantic meaning (e.g., *red butterfly* as the name of a butterfly species). For instance, in the task Game Night (see Figure 2), different kinds of chips labeled *blue*, *green*, or *red chips* serve as input variables, whereas different kinds of playing cards labeled *Royal*, *Grande*, or *Nobilis* serve as output variables.

While working on a MicroDYN task, a problem solver faces two different phases. In Phase 1, problem solvers can freely explore the system by entering values for the input variables (e.g., varying the amount of blue, green, and red chips in Figure 2). This is considered an evaluation-free exploration, which allows problem solvers to engage with the system and to use their knowledge acquisition ability under standardized conditions without controlling the system (Kröner et al., 2005). During Phase 1, problem solvers are asked to draw the connections between variables onscreen (see bottom of Figure 2), thereby producing data reflecting the knowledge acquired (3 min for Phase 1). Mayer (2003) calls this a situational external representation of a problem. In this first phase, the amount and correctness of explicit knowledge gathered during exploration are measured and expressed in a mental model as the final external problem representation (Funke, 2001). In Phase 2, problem solvers are asked to reach given target values on the output variables (e.g., card piles Royal, Grande, and Nobilis in Figure 2) by entering correct values for the input variables, thereby producing data reflecting the application of their knowledge (1.5 min for Phase 2). In this second phase, the goal-oriented use of knowledge is assessed.

These two phases are directly linked to the concepts of knowledge acquisition (i.e., representation) and knowledge application (i.e., generating and acting out a solution; Novick & Bassok, 2005). More detailed information on both the underlying formalism and the MicroDYN approach can be found in Funke (2001); Greiff et al. (2012), and Wüstenberg et al. (2012).

Purpose of Study and Hypotheses

```
graph LR; A[A] --> X[X]; B[B] --> Y[Y]; B[B] --> Z[Z]; C[C] --> Z[Z]; X[X] --> X[X]; Y[Y] --> Z[Z];
```

Input Variables

Output Variables

Figure 1. Structure of a typical MicroDYN task displaying three input (A, B, C) and three output (X, Y, Z) variables.

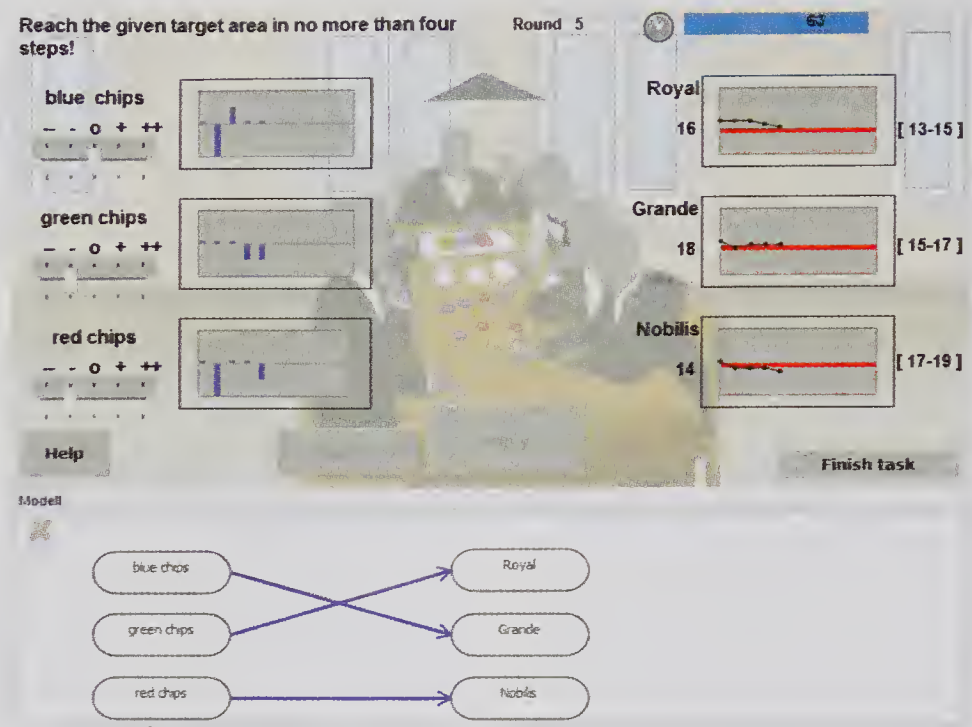


Figure 2. Screenshot of the MicroDYN task “Game Night.” The controllers of the input variables range from “– –” (value = –2) to “++” (value = +2). The current values and the target values of the output variables are displayed numerically (e.g., current value for Royal: 16; target values: 13–15) and graphically (current value: dots; target value: red line). The correct model is shown at the bottom of the figure.

knowledge acquisition); (2) the structural stability of the CPS construct across different grade levels of high school students ages 11 to 17 years; (3) latent mean comparisons between these grade levels if measurement invariance was sufficiently met; and (4) structural relations between CPS, fluid intelligence as a proximal measure of *g*, grade point average (GPA), and parental education across these groups to assess construct validity.

With regard to (1) the dimensionality of CPS, a large body of conceptual literature has suggested that the two CPS processes, knowledge acquisition and knowledge application, are related and yet somewhat distinct aspects of an overarching CPS process, but empirically this has been shown only for very selective samples (e.g., Kröner et al., 2005; Wüstenberg et al., 2012) and not yet for high school students. As part of assessing structural validity, we adhered to the question of whether there is an adequate construct representation of CPS by testing a measurement model that was closely aligned with the idea of partially separate mechanisms for problem representation and problem solution and assumed a two-dimensional model composed of the dimensions knowledge acquisition and knowledge application.

Hypothesis 1: We expected CPS to be composed of two different processes, knowledge application and knowledge acquisition. Thus, a two-dimensional model was expected to show the best fit and to fit significantly better than a one-dimensional model with the two processes combined under one first-order factor.

The structural stability of CPS, (2), pertains to the exact nature of the construct assessed. That is, the structure of the construct was not expected to change across different grade levels, indicating that the interpretation of test scores does not depend on the specific group the test is administered to (Byrne & Stewart, 2006). This

was tested by evaluating measurement invariance. Only to the extent that measurement invariance exists are between-group differences of grade levels unambiguous and able to be interpreted as true and not as psychometric differences in latent ability (Cheung & Rensvold, 2002, cf. Results section). For instance, it may be that due to cognitive development that occurs during adolescence, the construct of CPS changes. Analyses of measurement invariance would show that tasks behave differently in groups in different grade levels just as self-ratings on questionnaires may change their meaning when questions are translated from one culture to another (F. F. Chen, 2008).

Hypothesis 2: We expected CPS to show measurement invariance across different school grades.

The aspect of measurement invariance led directly to (3) latent mean comparisons of different grade levels or, in other words, to the question of level stability. For those parts of the measurement model that are identified as invariant, latent factor means can be compared, thus providing important insights into the effects of schooling and environment on CPS.

Hypothesis 3: If measurement invariance was sufficiently met, we expected latent mean differences between groups to indicate that students in higher grades perform significantly better in knowledge acquisition and knowledge application than students in lower grades.

In addition to establishing the validity of the internal structure, another important step is (4) establishing construct validity in terms of divergent and convergent relations to other constructs. To this end, we assessed how CPS was related to a measure of *g*, GPA, and parental education. Whereas GPA is an excellent marker of academic achievement, parental education reflects one of the

most important socioeconomic variables with a strong impact on school performance and educational outcomes (Myrberg & Rosen, 2008).

Hypothesis 4: Concerning construct validity, we expected (a) that *g* would predict performance on CPS tasks. However, a considerable amount of CPS variance was expected to remain unexplained, suggesting that parts of CPS are independent from *g* and (b) that CPS would predict GPA beyond *g*, as indicated by conceptual considerations and previous research. Furthermore, we expected (c) that parental education would predict performance in CPS and in *g*.

The field of CPS lags behind the field of intelligence testing, in which a broad range of well-established and extensively validated assessment procedures exist, some of which are even specifically tailored to educational demands (e.g., Wechsler, 2008). Considering the current educational interest in the assessment of CPS and the associated implications for researchers as well as practitioners such as educators and policymakers, this is particularly troublesome. By addressing the four research questions above, we aimed to make the measurement of CPS more evidence-based, thereby helping the field of CPS to catch up.

Method

Participants

Our sample ($N = 855$) was a subsample of a larger and more representative sample ($N > 4,000$) from a study conducted in Hungary. Participants were randomly drawn from Grades 5 to 11 in Hungarian elementary schools (Grades 5 to 8) and secondary schools (Grades 9 to 12).

Some software problems occurred during online testing, resulting in data loss. However, data were missing completely at random. Participants who were missing more than 50% of their data on MicroDYN or any other measure were excluded from all analyses (only about 5% of participants provided less than 80% data); other missing data were excluded on a pairwise basis.

Finally, data from 855 students were available for the analyses of Hypotheses 1 to 3 (mean age = 14.11 years, $SD = 1.83$; 46% boys). However, all analyses including those involving *g* (Hypotheses 4a and 4b) were based on a smaller subsample of students who completed both tests of CPS and *g* ($N = 486$; mean age = 14.36 years, $SD = 1.16$; 45% boys). Data were missing by design because *g* was not assessed in Grades 5, 6, and 11, and only a small number of missing values occurred due to drop-out (e.g., illness of students).

Design

CPS. MicroDYN was administered on computers. At the beginning, participants were instructed how to complete a trial task, in which they learned how the interface of the program could be controlled and which two tasks they were expected to solve: Participants explored unknown systems and drew their conclusions about how variables were interconnected in a situational model (cf. bottom of Figure 2; Mayer, 2003). This situational model was seen as an appropriate way of representing gathered information and allowed participants to visualize their mental model (knowledge

acquisition; Funke, 2001). Subsequently, they controlled the system by reaching given target values (knowledge application). After having finished the instruction phase, participants were given eight consecutive MicroDYN tasks. One task had to be excluded from analyses due to low communality ($r^2 = .03$) caused by an extreme item difficulty on knowledge acquisition ($p = .03$). All subsequent analyses were based on seven tasks. The task characteristics of all tasks (e.g., number of effects) were varied to produce tasks with an appropriate difficulty for high school students (cf. Greiff et al., 2012; see Appendix for equations).

g. The Culture Fair Test 20-R (CFT) consists of four subscales that measure fluid intelligence, which is seen as an excellent marker of *g* (Weiß, 2006) and is assumed to be at the core of intelligence (Carroll, 2003).

Dependent Variables and Scoring

CPS. Both MicroDYN dimensions, knowledge acquisition and knowledge application, were scored dichotomously, which is an appropriate way to score CPS performance (see Greiff et al., 2012; Kröner et al., 2005; Wüstenberg et al., 2012). For knowledge application, users' models were evaluated and credit was given for a completely correct model, whereas no credit was given when a model contained at least one mistake. Knowledge application was scored as correct when all target values of the output variables were reached.

g. All items of the CFT were scored dichotomously according to the recommendations in the manual (Weiß, 2006).

GPA and parental education. Participants self-reported their GPA from the previous school year and the educational levels of their parents. GPA ranged from 1 (*insufficient*) to 5 (*best performance*). Parental educational level for both mothers and fathers was scored on an ordinal scale (1 = *no elementary school graduation*; 2 = *elementary school*; 3 = *secondary school*; 4 = *university-entrance diploma*; 5 = *lower level university*; 6 = *normal university*; 7 = *PhD*).

Procedure

Test execution took place in the computer rooms of the participating Hungarian schools and lasted approximately 90 min. Participants worked on MicroDYN first, and the CFT was administered afterwards. Finally, participants provided demographic information. MicroDYN was delivered through the online platform *Testing Assisté par Ordinateur* (computer-based testing). Testing sessions were supervised either by research assistants or by teachers who had been trained in test administration.

Results

Descriptive Statistics

Analyses of manifest variables showed that the internal consistencies of MicroDYN as measures of CPS were acceptable (knowledge acquisition: $\alpha = .75$; knowledge application: $\alpha = .74$) and Cronbach's α for the CFT ($\alpha = .88$) was good. Participants' raw score distributions on the CFT ($M_7 = 39.84$, $SD = 9.13$; $M_8 = 41.36$, $SD = 7.54$; $M_9 = 36.97$, $SD = 7.20$; $M_{10} = 38.37$, $SD = 8.02$) differed slightly compared to the

original scaling sample of students attending the same grades ($M_7 = 34.98, SD = 6.63; M_8 = 36.37, SD = 6.56; M_9 = 38.42, SD = 6.43; M_{10} = 39.31, SD = 6.90$; Weiß, 2006). Further, participants' GPA showed a sufficient range ($M_7 = 4.00, SD = 0.80; M_8 = 3.95, SD = 0.83; M_9 = 3.64, SD = 1.05; M_{10} = 3.77, SD = 0.74; M_{11} = 3.64, SD = 0.71; 1 = insufficient, 5 = best performance$), and so did mothers' and fathers' education scores ($M_{mother} = 3.85, SD = 1.09; M_{father} = 3.75, SD = 1.10; 1 = no elementary school graduation, 7 = PhD$).

Statistical Analyses and Data Transformation

The analyses on the dimensionality of CPS (Hypothesis 1), measurement invariance (Hypothesis 2), latent mean differences (Hypothesis 3), and construct validity including only CPS and *g* (Hypothesis 4a) were based on latent models using structural equation modeling (SEM; Bollen, 1989). SEM analyses using latent variables require larger sample sizes than traditional statistics based on manifest variables. On this matter, Ullman (2007) recommended that the number of estimated parameters should be no more than one fifth of *N*. To meet this guideline, we merged Grades 5 and 6, Grades 7 and 8, as well as Grades 10 and 11, to Grade Levels 5/6, 7/8, and 10/11, respectively, so that sufficient data were provided to test measurement models separately within each group or grade level, respectively. We kept Grade 9 as a single grade level because the transition from elementary to secondary school takes place after Grade 8 in the Hungarian school system. This transition is known to affect cognitive performance and to be associated with a general loss in achievement (e.g., Alspaugh & Harting, 1995; S. S. Smith, 2006). Specifically, Molnár and Csapó (2007) reported a drop in problem-solving performance in Grade 9 test scores in Hungary. Even though we did not pose any hypotheses about the performance pattern in Grade 9, we did not merge these students in order to be able to detect effects of the transition. All other analyses including GPA, CPS, *g*, and parental education (Hypotheses 4b and 4c) were based on manifest (observed) data (cf. results on Hypotheses 4b and 4c). Mplus 5.0 was used for all analyses (Muthén & Muthén, 2010).

Hypothesis 1: Dimensionality of CPS

We used confirmatory factor analyses within SEM to test the underlying measurement model of CPS with the two different CPS processes knowledge acquisition and knowledge application (Hypothesis 1). Table 1 shows the dimensionality results. The two-

dimensional model fit well in the overall sample compared to cut-off values recommended by Hu and Bentler (1999), who stated that comparative fit index (CFI) and Tucker Lewis index (TLI) values above .95 and a root mean square error of approximation (RMSEA) below .06 indicate a good global model fit. Within the two-dimensional model, the measures of knowledge acquisition and application were significantly correlated on a latent level ($r = .74, p < .001$; manifest correlation: $r = .52, p > .001$). When estimating this and all subsequent models, we used the preferred estimator for categorical variables: the weighted least squares mean and variance adjusted estimator (WLSMV; Muthén & Muthén, 2010).

We also tested a one-dimensional model with all indicators combined under one general factor; however, the fit indices decreased considerably. In order to compare the two-dimensional and one-dimensional models, χ^2 values in Table 1 cannot be directly subtracted to compare them because computing the differences of χ^2 values and *dfs* between models is not appropriate if WLSMV estimation is applied (Muthén & Muthén, 2010, p. 435). Thus, we carried out a χ^2 difference test in Mplus (Muthén & Muthén, 2010), which showed that the two-dimensional model fit significantly better than the one-dimensional model ($\chi^2 = 86.121, df = 1, p < .001$). After this, the two-dimensional model was applied to each grade level (i.e., Grade Levels 5/6, 7/8, 9, and 10/11) separately, also showing a very good fit (see Table 1).

In summary, the two-dimensional model fit well in the overall sample and for each grade level. Thus, the processes knowledge acquisition and knowledge application were empirically distinguished, supporting Hypothesis 1.

Measurement Model of *g*

As a prerequisite for all analyses involving *g*, we had to test a measurement model for the CFT. Because the CFT contains 56 items, we decided to use the item-to-construct balance recommended by Little, Cunningham, Shahar, and Widaman (2002) to assign items to four parcels. Each parcel consisted of 14 CFT items to reduce the number of parameters to be estimated. The mean difficulty of the parcels did not differ significantly ($M_1 = .72; M_2 = .75; M_3 = .71; M_4 = .66; F(3, 56) = 0.52, p > .05$) and the parcels' factor loadings were also comparable ($\beta_1 = .82, \beta_1 = .78, \beta_1 = .80, \beta_1 = .78; F(3, 56) = 0.33; p > .05$). The measurement model with *g* based on four parcels showed a very good fit for the overall sample ($N = 486; \chi^2 = .717; df = 2; p > .05; CFI = .999; TLI = .999$;

Table 1
Goodness of Fit Indices for Testing Dimensionality of MicroDYN, Overall and by Grade Level

Model	χ^2	<i>df</i>	<i>p</i>	CFI	TLI	RMSEA	<i>n</i>
Two-dimensional including all grade levels	164.068	53	.001	.967	.978	.050	855
One-dimensional including all grade levels	329.352	52	.001	.912	.944	.079	855
Two-dimensional, Grade Level 5/6 only	65.912	35	.001	.966	.966	.064	216
Two-dimensional, Grade Level 7/8 only	77.539	13	.001	.969	.969	.056	300
Two-dimensional, Grade Level 9 only	13.908	29	.380	.996	.996	.029	83
Two-dimensional, Grade Level 10/11 only	51.338	40	.001	.991	.991	.033	256

Note. χ^2 and *df* were estimated by the weighted least squares mean and variance adjusted estimator (WLSMV). CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation.

RMSEA = .001), as well as for the different grade levels (CFIs = .991–.999; TLIs = .991–.999; RMSEAs = .001–.002).

Hypothesis 2: Measurement Invariance

Measurement invariance was tested by multigroup analyses using the means and covariance structure (MACS) approach within SEM. The general procedure of testing measurement invariance is explained in detail by Byrne and Stewart (2006). They describe a series of hierarchical steps that have to be carried out such that each step imposes an increasingly greater number of restrictions to model parameters to test invariance. Thereby, four different models of invariance are distinguished: configural invariance, weak factorial invariance, strong factorial invariance, and strict factorial invariance. In general, measurement invariance is met if restrictions of model parameters in one model do not generate a substantially worse model fit in comparison to an unrestricted model. The model fit can be evaluated by either a *practical* perspective, reflected in a drop in fit indices such as the CFI (CFI < .01; Cheung & Rensvold, 2002), or by a *stricter traditional* approach, indicated by a significant χ^2 difference test. Only if at least strong factorial invariance is established can latent mean comparisons (Hypothesis 3) be meaningfully interpreted. Otherwise, between-group differences may reflect psychometric properties of the items and not true differences (Byrne & Stewart, 2006).

CPS. To test the measurement invariance of MicroDYN, we applied a procedure that is slightly different from the typical one recommended by Byrne and Stewart (2006). MicroDYN data were based on categorical variables, and thus constraints on model parameters differed in comparison to invariance tests based on continuous variables (Muthén & Muthén, 2010, p. 433).

Indices of global model fit for all analyses on measurement invariance are shown in Table 2. Based on the two-dimensional model derived in Hypothesis 1, this multigroup model testing configural invariance of CPS fit well. In this model, thresholds² and factor loadings were not constrained across groups, factor means were fixed to zero in all groups, and residual variances were fixed to one in all groups (as recommended by Muthén & Muthén, 2010, p. 434) instead of freely estimating residuals as is done with continuous outcomes. Weak factorial invariance was not tested because it is not recommended when the WLSMV estimator for categorical outcomes is used (Muthén & Muthén, 2010, p. 433). Thus, the next step was to test for strong factorial invariance, in which thresholds and factor loadings were constrained to be equal across groups, residual variances were fixed to one, and factor means were fixed to zero in one group (i.e., Grade Level 5/6), whereas there were no constraints specified in any other group (Muthén & Muthén, 2010, p. 343). The strong factorial invariance model did not show a decrease in model fit based on the practical perspective ($\Delta\text{CFI} < .01$) or based on the stricter traditional perspective (nonsignificant χ^2 difference test; see Table 2) compared to the configural invariance model. Finally, we evaluated strict factorial invariance, in which, in addition to the restrictions realized in strong factorial invariance, all residual variances were fixed to one in all groups. Results from Table 2 showed that MicroDYN was also invariant in a strict sense, even though strict factorial invariance is not a prerequisite for group comparisons of latent factor means and variances (see Byrne & Stewart, 2006).

Although invariance was found for MicroDYN, suggesting an identical factor structure across grade levels, single path coefficients can differ without compromising the invariance of the overall model. This would account for correlations between measures of knowledge acquisition and knowledge application, which varied across the different grade levels ($r_{5/6} = .82$, $SE = .05$; $r_{7/8} = .68$, $SE = .05$; $r_9 = .94$, $SE = .06$; $r_{10/11} = .72$, $SE = .05$). The two dimensions correlated significantly higher in Grade Level 9 than in Grade Level 5/6 (based on z statistics), which in turn showed a significantly higher correlation than Grade Levels 10/11 and 7/8, whereas the latter two did not differ significantly (Grade Level 10/11 = Grade Level 7/8 < Grade Level 5/6 < Grade Level 9). These findings raised some concerns about the pattern of results for Grade 9; these are discussed later on in more detail.

In summary, MicroDYN showed measurement invariance so that latent mean differences could be interpreted as true differences in the construct being measured (Byrne & Stewart, 2006). Consequently, Hypothesis 2 was supported.

g. As a prerequisite for Hypothesis 4, we tested for construct validity between CPS, *g*, and external criteria. At this stage, we also checked for the measurement invariance of the CFT as described in the Method section (and as recommended by Byrne & Stewart, 2006). We used maximum likelihood estimation for continuous variables for all models because CFT data were parceled and could be considered continuous. The CFT was invariant in a strict sense as indicated by a nonsignificant χ^2 difference ($p > .10$) between the models of strict factorial invariance ($\chi^2 = 25.546$, $df = 26$; CFI = .999, TLI = .999, RMSEA = .001) and configural invariance ($\chi^2 = 3.908$, $df = 6$; CFI = .999, TLI = .999, RMSEA = .001).

Hypothesis 3: Latent Mean Comparisons

CPS. As a prerequisite for comparing means across groups, the MicroDYN scale had to be fixed to a user-specified level by setting the latent means of a reference group to zero in both dimensions (e. g., Grade Level 5/6), whereas the latent means of all other groups were freely estimated and subsequently compared to the reference group. Thus, we used the strong factorial invariance model and compared all grade levels with each other, starting with Grade Level 5/6 as the reference group (left part of Table 3), whereas Grade Level 7/8 served as the reference group in a second comparison (middle part of Table 3) and Grade Level 9 in a third comparison. It was expected that all latent means would have a positive value and would differ significantly from the corresponding reference groups, thereby indicating that students in higher grade levels performed better.

Results for measures of knowledge acquisition indicated that Grade Level 9 performed worse than Grade Level 5/6 (cf. Table 3), which in turn performed worse than Grade Levels 7/8 and 10/11, whereas the means of the latter two grade levels did not differ significantly (rank order: Grade Level 9 < Grade Level 5/6 < Grade Level 7/8 = Grade Level 10/11). Comparisons between the latent means of the measures of knowledge application scores showed that, once again, Grade Level 9 performed the worst, followed by Grade Levels 5/6 and 10/11, neither of which differed

² In models containing categorical variables, thresholds are used instead of intercepts.

Table 2
Goodness of Fit Indices for Measurement Invariance of MicroDYN

Model	χ^2	df	Compared with	$\Delta\chi^2$	Δdf	p	CFI	TLI	RMSEA
(1) Configural invariance	161.045	104					.975	.975	.051
(2) Strong factorial invariance	170.101	115	(1)	22.294	23	>.10	.976	.982	.047
(3) Strict factorial invariance	165.826	116	(1)	53.159	43	>.10	.978	.983	.045

Note. χ^2 and *df* were estimated by the weighted least squares mean and variance adjusted estimator (WLSMV). $\Delta\chi^2$ and Δdf were estimated by the Difference Test procedure in MPlus (see Muthén & Muthén, 2010). Chi-square differences between models cannot be compared by subtracting χ^2 s and *dfs* if WLSMV estimators are used. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation.

significantly from Grade Level 9. Grade Level 7/8 performed better than all other grade levels (rank order: Grade Level 9 < Grade Level 5/6 = Grade Level 10/11 < Grade Level 7/8).

g. Similar to MicroDYN, Grade Level 9 had significantly lower means on CFT scores compared to Grade Level 7/8 ($M_{7/8} = 0$; $M_9 = -.55$, $SE = .15$, $p < .01$) and also compared to Grade Level 10 ($M_{10} = 0$; $M_9 = -.38$, $SE = .17$, $p < .05$), whereas the latter did not differ significantly from Grade Level 7/8 ($M_{7/8} = 0$; $M_{10} = -.15$, $SE = .12$, $p > .05$). The overall order of the means was comparable to the pattern for measures of knowledge acquisition (rank order: Grade Level 9 < Grade Level 7/8 = Grade Level 10; the CFT was not administered to Grades 5, 6, and 11).

In summary, findings were not as straightforward as expected because performance on all measures did not increase consistently in higher grade levels. In addition to the generally low performance in Grade Level 9 on all measures, measures of knowledge application scores dropped for Grade Level 10/11 compared to Grade Level 7/8, whereas measures of knowledge acquisition remained stable. Thus, Hypothesis 3 was only partially supported.

Hypothesis 4: Construct Validity

All analyses to test relations between CPS and *g* (Hypothesis 4a) used models with latent variables within structural equation modeling. However, results for CPS, *g*, GPA, and parental education (Hypotheses 4b and 4c) were based on path analyses using manifest variables because the sample sizes of the subsamples (e.g., Hypothesis 4b: $N = 75$ in Grade 11) were not appropriate for latent analyses.

CPS and *g*. We assumed that *g* would predict CPS performance; however, a significant amount of variance was expected to remain unexplained (Hypothesis 4a). Thus, by using structural equation modeling, we regressed MicroDYN on the CFT and estimated the proportion of explained variance in the MicroDYN dimensions. The results, illustrated in Table 4, showed that the CFT explained performance in measures of knowledge acquisition and knowledge application in the overall model, as well as in all separate grade level models. Although the CFT significantly predicted performance for both dimensions, the residuals of measures of knowledge acquisition and knowledge application were still highly correlated ($rs = .32-.62$), indicating common aspects of CPS dimensions separable from *g*. The model fit well for the overall sample (CFI = .948, TLI = .971, RMSEA = .053) and showed a good to acceptable fit for the several grade level models (CFIs = .932–.992, TLIs = .960–.994, RMSEAs = .032–.062). Except for Grade Level 9 ($p < .01$), path coefficients of the CFT predicting the dimensions acquisition and application (left part of Table 4) differed only marginally between grade levels ($p > .05$).

Overall, participants in Grade Level 9 showed unexpected data patterns for Hypotheses 2, 3, and 4a: They scored the worst by far on MicroDYN and the CFT, in comparison to both other grade levels and the CFT scaling sample. Further, measures of knowledge acquisition and knowledge application were extremely highly correlated in Grade Level 9 (see results in Hypothesis 2). Also, MicroDYN and the CFT were related more strongly than in all other grade levels (see Hypothesis 4a and residual correlations in Table 4). The combination of poor performance on all measures

Table 3
Latent Mean Comparisons of Knowledge Acquisition and Knowledge Application (MicroDYN) Between Different Grade Levels

Model	Compared with (1)			Compared with (2)			Compared with (3)		
	<i>M</i>	<i>SE</i>	<i>p</i>	<i>M</i>	<i>SE</i>	<i>p</i>	<i>M</i>	<i>SE</i>	<i>p</i>
Acquisition									
(1) Grade Level 5/6		.00							
(2) Grade Level 7/8	.18	.11	<.05						
(3) Grade Level 9	–.37	.17	<.05	–.54	.15	<.001			
(4) Grade Level 10/11	.30	.15	<.05	.04	.13	>.05	.88	.36	<.01
Application									
(1) Grade Level 5/6		.00							
(2) Grade Level 7/8	.50	.24	<.05						
(3) Grade Level 9	–.52	.25	<.05	–.72	.29	<.001			
(4) Grade Level 10/11	.04	.13	>.05	–.24	.11	<.05	.88	.36	<.01

Note. Statistical significance of the differences between all groups was determined by *z* statistics.

Table 4
Prediction of Performance in Knowledge Acquisition and Knowledge Application (MicroDYN) by g, Overall and by Grade Level

Model	Path coefficient		R^2		Residual correlation acquisition/application	n
	Acquisition	Application	Acquisition	Application		
Overall	.47*** (.04)	.40*** (.05)	.22*** (.04)	.16*** (.04)	.63*** (.05)	486
Grade Level 7/8	.48*** (.05)	.39*** (.07)	.23*** (.05)	.15*** (.05)	.60*** (.06)	284
Grade Level 9	.62*** (.12)	.62*** (.12)	.38*** (.14)	.38*** (.15)	.30*** (.10)	79
Grade Level 10	.34*** (.10)	.32*** (.11)	.11* (.07)	.11* (.07)	.62*** (.08)	123

Note. Standard errors are in parentheses.
* $p < .05$. ** $p < .01$. *** $p < .001$.

and increased correlations between them indicate that covariates strongly influenced performance scores. Thus, we decided to elaborate on possible reasons for the unexpected pattern of results in the Discussion section and to exclude Grade Level 9 from further analyses.

CPS, g, and GPA. Having shown that MicroDYN had a significant amount of unshared variance with the CFT, we thought it possible that the two constructs might also differ in their predictive validity, further indicating that CPS is separable from g. Thus, we checked the incremental validity of MicroDYN beyond the CFT in predicting performance in GPA (Hypothesis 4b). We decided to use grades (e.g., Grades 7 and 8, separately) instead of grade levels (e.g., Grade Level 7/8) in these analyses because school GPA is not comparable between different grades, and the same GPA in different grades (e.g., Grade 7 or Grade 8) reflects different levels of performance.

Whereas scores on MicroDYN and the CFT were based on the same test for all students, GPA depended on demands that varied across grades. We used manifest path analyses due to the small sample sizes within each grade: As shown in Table 5, the criterion GPA was predicted by only MicroDYN, only CFT, and, in a final step, by MicroDYN and the CFT simultaneously. In the last model, both predictors, the CFT and MicroDYN, were combined to determine the incremental validity of MicroDYN by comparing the explained variance of this model with the explained variance of the model containing only the CFT (indicated by ΔR^2 in Table 5).

Table 5
Prediction of GPA by MicroDYN and CFT

Grade	R^2 in GPA			ΔR^2	n
	MicroDYN	CFT	MicroDYN and CFT		
7	.03*	.19***	.19***	.00	104
8	.08*	.09***	.13***	.04*	93
10	.07*	.15***	.18***	.03*	90
11	.07*				75

Note. R^2 = explained variance. Significant ΔR^2 s indicate significant path coefficients of CPS contributing to R^2 . GPA = grade point average; CFT = Culture Fair Test 20-R; CPS = complex problem solving.
* $p < .05$. *** $p < .001$.

Results displayed in Table 5 show that although MicroDYN predicted performance in GPA, the CFT was more strongly related to GPA. Additionally, MicroDYN added a small percentage of variance when predicting GPA together with the CFT in Grades 8 and 10. Global model fit was good (RMSEAs = .000–.001, CFIs = .991–.999). Thus, Hypothesis 4b was supported even though this finding was not consistent across all grades.

CPS, g, and parental education. To investigate the impact of potential determinants of CPS, we hypothesized that parental education would predict performance for MicroDYN and the CFT (Hypothesis 4c). We used path analysis because of the small sample sizes within each grade and predicted performance in MicroDYN and the CFT by parental education. Results showed that mothers' education predicted performance in MicroDYN in Grade 7 ($R^2_{\text{MicroDYN}} = .03, p < .05; R^2_{\text{CFT}} = .00, p > .05$) and Grade 8 ($R^2_{\text{MicroDYN}} = .06, p < .05; R^2_{\text{CFT}} = .03, p > .05$) but not performance on the CFT. The opposite was true in Grade 10 ($R^2_{\text{MicroDYN}} = .00, p > .05; R^2_{\text{CFT}} = .04, p < .05$). Fathers' education yielded significant paths for MicroDYN and the CFT only in Grade 7 ($R^2_{\text{MicroDYN}} = .02, p < .05; R^2_{\text{CFT}} = .02, p < .05$), although fathers' education was significantly correlated with mothers' education ($r = .54, p < .01$). In summary, mothers' education predicted performance in MicroDYN and on the CFT, even though this finding was not consistent across all grades, partially supporting Hypothesis 4c.

Discussion

The aim of the present study was to enhance the understanding of complex problem solving and to evaluate its relevance in educational contexts by defining the concept and by establishing construct validity in a sample of Hungarian high school students. Generally, the results of the current study provided support for an understanding of CPS as a broad mental process measurable by means of computer-based assessment with high relevance to education. More specifically, (a) CPS was best modeled as a two-dimensional construct with the dimensions knowledge acquisition and knowledge application, (b) measurement of these two dimensions was invariant across groups composed of Hungarian high school students ranging from 11 to 17 years in age, and (c) latent mean comparisons revealed an increase in knowledge acquisition and in knowledge application in part (i.e., only from Grade Level 5/6 to Grade Level 7/8) with increasing grade level. However, this was not true for students in Grade 9, who performed the lowest on both dimensions, as we discuss later on. (d) CPS was correlated

with and yet clearly distinct from a measure of *g* and exhibited predictive validity beyond it. Further, level of parental education was related to CPS and *g*, yielding overall important educational implications for the understanding of complex cognitive abilities such as CPS.

Dimensionality: Knowledge Acquisition and Knowledge Application

The data showed the best fit to the model that assumed the existence of two dimensions of CPS, knowledge acquisition and knowledge application. This finding supports a common assumption that knowledge acquisition is a necessary but not a sufficient condition for knowledge application. For instance, Newell and Simon (1972) stated that goal-oriented problem solving necessitates an adequate problem space in which important knowledge about the problem is stored. However, they also acknowledged that generating and applying a solution depends on additional procedural abilities, such as forecasting, strategic planning, or carrying out planned actions (Raven, 2000). Consequently, research on CPS has generally applied a knowledge acquisition and a subsequent knowledge application phase (e.g., Kröner et al., 2005). Results in this study supported these findings within a psychometric assessment approach for different grade levels of students.

Usually, ability assessment is limited to the evaluation of final solutions. That is, the final results of cognitive processes, for instance, knowledge application scores in CPS, are used in educational contexts to make selection decisions, to initiate specific training measures, or to assess an entire educational system. However, the cognitive process of deriving a representation and actually carrying out a problem solution is often disregarded, but some added value is to be expected by establishing more process-oriented measures. Clearly, CPS with its broad components is a valid candidate for such an enterprise, and future research should attend to the issue of process measures as their added value becomes available through computer-based assessment.

Measurement Invariance Across Grade Levels (Structural Stability)

Comparing CPS scores between grade levels requires that the assessment instrument, MicroDYN, measure exactly the same construct across groups as indicated by measurement invariance. The current study tested CPS for strong invariance of a first-order structure composed of the two dimensions knowledge acquisition and knowledge application. According to Byrne and Stewart (2006), evidence of invariance can be based on either a traditional perspective by evaluating significant drops in overall fit or on a more practical perspective by evaluating absolute changes in fit criteria. As portrayed in Table 2, results from either perspective strongly supported the invariance of CPS in Hungarian students across Grade Levels 5 through 11, which generally speaks well for the MicroDYN measure and its adoption in Hungary. That is, individual differences in factor scores are due to differences in underlying ability, allowing direct comparisons of ability levels between students and between grades.

Results of tests of measurement invariance can also provide insight into the structural development and the structural stability of knowledge acquisition and knowledge application even though

these are somewhat limited by the cross-sectional nature of the data. Whereas no studies have addressed the issue of structural stability in CPS until now, much is known about it in *g*. A large body of studies has suggested that both *g* on Stratum III and broad cognitive abilities on Stratum II within the CHC theory are shaped by the time students begin attending school (e.g., Salthouse & Davis, 2006). That is, the factorial structure of *g* is built early in childhood (no later than by the age of 6) and then remains constant for several decades. It is only in older age that differentiation may once again decrease, as indicated by increasing correlations among Stratum II abilities and higher factor loadings on *g* (Deary, Whalley, & Crawford, 2004). CPS is composed of complex mental operations (Funke, 2010). Thus, differentiation into knowledge acquisition and knowledge application is unlikely to take place earlier than it takes place in *g*. As strict factorial invariance holds from Grade Level 5/6 (youngest age: 11 years) to Grade Level 10/11, this differentiation cannot take place before the age of 6 but has largely taken place by the age of 11. That is, the results of our study suggest that at the age of 11, the structural stability of CPS can be assumed.

Latent Mean Comparisons Across Grade Levels

After finding evidence of an invariant factor structure, the study tested latent mean differences between grade levels. Results revealed that the mean scores of Grade Level 7/8 were higher than those of Grade Level 5/6, whereas Grade Level 9 scored the lowest on both indicators. Grade Level 10/11 showed the same performance as Grade Level 7/8 in knowledge acquisition but showed a small and yet significant decrease in knowledge application.

Not entirely unexpected was that latent scores of students in Grade Level 9 exhibited a substantial drop in performance on both dimensions and, additionally, on latent scores of the CFT. This drop and the consolidation of performance in Grade Level 10/11 can be seen in the context of the transition from elementary to secondary school in Hungary, which takes place just before entering Grade 9. School transitions in general yield personal and academic challenges and are highly likely to be associated with achievement loss (e.g., S. S. Smith, 2006). In the specific case of Hungary, Molnár and Csapó (2007) also reported a general decrease in test scores in Grade 9 for Hungarian students, thus showing that this performance decrease is not limited to our sample. These drops in academic performance tend to recover to their pretransitional levels in the year following the transition (Alspaugh & Harting, 1995).

There is a mutual understanding among researchers that transition impairs achievement. However, little is known about the underlying mechanisms. Besides stress imposed by the distracting nature of changing peer relationships, new norms, and harsher grading compared to elementary school (Alspaugh & Harting, 1995), a general loss of motivation partly attributable to effects of pubertal changes (Wigfield, Byrnes, & Eccles, 2006) is assumed to further attenuate test performance (S. S. Smith, 2006). In our study, not only was mean performance level higher, but latent correlations between knowledge application, knowledge acquisition, and *g* were also strikingly higher in Grade 9 than in any other grade level, possibly pointing to motivational issues as the underlying cause. That is, as students were less motivated to perform well on any of the tests, the variance in performance scores was

largely generated by different levels of motivation, resulting in high correlations between constructs. This is a well-known effect in research on the development of intelligence. However, alternative explanations for the performance drop in Grade 9 are feasible as well. For instance, students in lower grades might have perceived the CPS task as some kind of game and enjoyed working on it, whereas tasks might have been simplistic and boring to students in higher grades.

Considering the significant drop precisely at the change from elementary to secondary school and the (partial) recovery in scores in Grade Level 10/11 at some point after the transition observed in our study, transition apparently plays a role in explaining performance patterns across grades. However, to reveal the underlying causes and to decide between competing explanations, more comprehensive and experimental studies are required. Therefore, we decided not to interpret the results from students in Grade Level 9 and to interpret results from Grade Level 10/11 with caution in all further analyses.

After we excluded Grade Level 9, a more consistent picture of latent means could be drawn. First, scores increased significantly from Grade Level 5/6 to Grade Level 7/8 for both CPS processes and *g*, showing a combined effect of school and out-of-school experiences, and even the literature acknowledges that schooling plays a large role in this development (Rutter & Maughan, 2002). Substantive interpretation of these results suggests that a change in mean scores may indeed reflect true between-grade-level differences, which is in line with research that has reported that substantial cognitive development takes place at this age (Byrnes, 2001).

However, the picture is different for the change in latent means from Grade Level 7/8 to Grade Level 10/11: Whereas *g* and knowledge acquisition remained at least stable, there was a statistically significant albeit small drop in performance for knowledge application. This is in contradiction to the work of Byrnes (2001), who claimed, without having studies including CPS available for his review, that both declarative knowledge and procedural knowledge increase with age during the adolescent period. Further, of the two CPS processes, the performance decrease in knowledge application from Grade Level 7/8 to Grade Level 10/11 was accompanied by decreasing latent correlations.³ That is, as knowledge acquisition and knowledge application exhibited different patterns of latent means across grade levels, they also became continuously less connected (shared variance dropped from 73% to 46%).

The potentially different developmental trajectories of knowledge acquisition and knowledge application and the change in correlation patterns in higher grades cannot be explained only as an effect of transition and its consequences because no drop from Grade Level 7/8 to Grade Level 10/11 was observed for knowledge acquisition, but rather only for knowledge application. Thus, there may be other causes that underlie this effect. This finding is in line with Spearman's (1927) law of diminishing returns, which claims that correlations between different tests decrease with increasing age, postulating a successive differentiation as time goes by. This conception has received considerable criticism from intelligence researchers but has not been considered for CPS. One possible explanation is that the development of knowledge application and knowledge acquisition may increasingly diverge across the life span, similar to what Spearman (1927) proposed for *g*, and as our data tentatively suggest.

Another explanation for the different development trajectories of knowledge acquisition and knowledge application is that the Hungarian school system is known as a traditional system with little emphasis on procedural knowledge as captured in knowledge application (Nagy, 2008). As a consequence, knowledge application skills might have deteriorated between Grade Levels 7/8 and 10/11, whereas knowledge acquisition and *g* were at least consolidated on a stable level. Clearly, these tentative results based on cross-sectional data have to be cautiously interpreted, and other interpretations may account equally well for the different development of the two dimensions knowledge acquisition and knowledge application. Thus, replications of these results are needed, as this is the first study on the development of CPS, but these findings point out interesting paths for future research.

Construct Validity: CPS, *g*, and External Variables

To shed further light on CPS and to relate it to other measures of cognitive performance, we investigated relations among CPS and *g*, GPA, and parental education. The most comprehensive and most widely acknowledged approach to understanding mental ability is found in the CHC theory, which assumes three hierarchically arranged strata of mental abilities with *g* located on a general Stratum III (McGrew, 2009). Two questions about CPS and CHC theory need to be answered: How does CPS relate to *g*? And how does CPS relate to the broad cognitive abilities on Stratum II?

Clearly, CPS is influenced by *g* (e.g., Kröner et al., 2005; Wüstenberg et al., 2012), but the path coefficients between *g* and CPS, which ranged from .32 to .62 in this study, were substantially lower than those usually reported between *g* and other Stratum II abilities. Does this imply that CPS cannot be subsumed within Stratum II? We did not explicitly measure Stratum II abilities, but we used the CFT to test fluid intelligence, which is assumed to be at the core of *g* (Carroll, 2003). In fact, fluid intelligence exhibits the highest factor loading on *g*, and some researchers suggest isomorphism between the two (e.g., Gustafsson, 1984). Considering that CPS is measured by dynamic and interactive tasks, whereas Stratum II abilities are exclusively measured by static tasks, which do not assess the ability to actively integrate information or to use dynamically given feedback to adjust behavior (Wüstenberg et al., 2012), CPS may indeed constitute one aspect of *g* that is not yet included within Stratum II. This may particularly hold for knowledge application, which exhibited lower correlations with *g* than did knowledge acquisition.

Sound measures of CPS have emerged only recently and were not available in studies that have tested the CHC theory. However, new Stratum II abilities, such as general knowledge or psychomotor speed, have been tentatively identified (McGrew, 2009) and have led to adaptations of the CHC theory. Further widening the view by including dynamic measures of CPS in future studies, as recently proposed by Wüstenberg et al. (2012), may turn out to increase the understanding of how mental ability is structured. Results in the current study, albeit tentative, suggest divergent validity between measures of *g* and CPS, even though the theoretical implications of these findings are not conclusive. On the

³ Please note that single latent correlations may differ without compromising strong measurement invariance and do not contradict the finding of invariance (Byrne & Stewart, 2006).

other hand, if CPS is really important and contributes to the explanation of students' performance in educational contexts, this should be reflected by the prediction of relevant external variables.

To test this assumption, we related g and CPS to GPA and checked whether CPS incrementally predicted GPA beyond g . We further related CPS to another relevant external variable, parental education. GPA is assumed to reflect the level of academic achievement over a longer period of time and was strongly related to g in our study. This is in alignment with a large body of research and is not surprising insofar as measures of g were originally constructed to predict academic performance in school (Jensen, 1998). In addition to g , representation of complex problems indicated by knowledge acquisition added a small percentage of explained variance, whereas the paths for knowledge application were mostly not substantial. Again, this was not surprising because the representation of acquired knowledge is demanded in school more frequently than is actively carrying out a pattern of solution steps (Lynch & Macbeth, 1998). Further, this pattern of results is in line with a recent study by Wüstenberg et al. (2012), who also reported the empirical significance of knowledge acquisition beyond measures of g in predicting GPA.

Parental education, which served as a predictor of both CPS and g in our study, has been shown to be the most important socioeconomic factor in influencing school performance (Myrberg & Rosen, 2008) and to be somewhat related to g . To this end, Rindermann, Flores-Mendoza, and Mansur-Alves (2010) reported a small yet significant relation of parental education and g . In our study, parental education predicted g as well as CPS, even though not consistently in all grades. One explanation for the significant relation between CPS and parental education, especially in earlier grades, may be that parents with higher levels of education provide more stimulating and activating learning environments, offer more emotional warmth, and often engage in playful and educational activities with their children (Davis-Kean, 2005). These children may be confronted more often with dynamic and interactive situations, which are fundamental for acquiring and applying new knowledge.

How can these findings further inform a theoretical understanding of g , CPS, and their reciprocal relation? Clearly, g is a good predictor of academic achievement, which can be somewhat complemented by CPS, as shown in this study and in Wüstenberg et al. (2012). Additional support for the relevance of CPS is found in Danner et al. (2011), who reported that CPS predicted supervisor ratings on the job beyond g . In summary, more research on the nature of CPS is needed to bolster the results found in this study, but the increase in the accuracy yielded by CPS in predicting relevant external criteria is a promising starting point.

Limitations

Obvious limitations of this study that require consideration refer primarily to sample characteristics and methodological issues: A cross-sectional design of a limited age span in only a few grade levels was used, thus prohibiting generalization of results and causal conclusions. Further, there might have been small flaws in the representativeness of our subsample, and these, paired with potentially influential transition effects, led to the exclusion of Grade 9 in the analyses on construct validity. We clearly acknowledge that relations between constructs may differ depending on the

methods applied (e.g., Myrberg & Rosen, 2008) and that, therefore, our results are to a certain extent tentative and not generalizable. However, a more severe problem that research on CPS suffers from is that few studies have addressed the issue of the assessment and construct validity of CPS. Thus, directly comparing our results to previous research is difficult, and interpretations remain inconclusive. Clearly, research will strongly benefit from widening the view to other designs.

A second point relates to the understanding of g in this study. By employing the CFT, we tested a rather narrow aspect of g , and it is difficult to relate CPS and the CHC theory when only single measures are applied. On the other hand, fluid intelligence is the strongest marker of g (Carroll, 2003) and one of its most frequently used tests. We suggest for further research to again widen the view by explicitly assessing different Stratum II abilities. However, just as our measure of g could be challenged, this is also true for the measure of CPS: The nature of the tasks we used heavily influenced the problem-solving process and narrowed it down to a certain extent, an issue faced by any latent construct. For instance, Newell and Simon (1972) suggested that problem solvers refer back to the problem space when carrying out a problem solution. This interaction between knowledge acquisition and knowledge application was not included in our study. On the other hand, the two main processes identified by problem-solving research (i.e., representation and solution) are theoretically implemented in our measure of CPS and were empirically separable. Further, careful attempts to develop CPS measures have been scarce until now, and our results suggest that using multiple complex tasks is a valid approach for capturing CPS performance.

Implications and Conclusion

The general impact of schooling on mental ability has been widely acknowledged (Rutter & Maughan, 2002). At the same time, enhancing cognitive performance in school or, in other words, improving students' minds is a major challenge of education and an educational goal in itself (Mayer & Wittrock, 2006). In fact, large-scale assessments such as PISA are explicitly aimed at describing and comparing levels of achievement in different educational systems, but the implicit goal is to find ways to make education more efficient, for example, by enhancing complex cognitions such as problem solving. When it comes to these complex cognitions, it is often assumed that this challenge is met implicitly in school. To describe this phenomenon, Mayer and Wittrock (2006) introduced the term *hidden curriculum*, stating that "educators expect students to be able to solve problems . . . but rarely provide problem-solving instruction" (p. 296). The assumption of a hidden curriculum may partly be unjustified, as the results of our study suggest: CPS and its components were not as strongly fostered as one might have hoped.

In the search for methods that promote CPS, Mayer and Wittrock (2006) listed seven instructional methods with a more or less proven impact on problem solving. However, one general disadvantage of approaches aimed at enhancing problem-solving skills is that evidence for transfer to other types of problems is rather scarce (Mansfield, Busse, & Krepelka, 1978). To this end, Mayer and Wittrock (2006) concluded that teaching domain-specific skills is more promising than trying to foster domain-general CPS abilities.

At this point, we are less pessimistic than and differ in our view from Mayer and Wittrock (2006). Similar to our position, Novick, Hurley, and Francis (1999) underline the importance of general processes in problem solving by stating that abstract representation schemas (e.g., causal models or concept maps) are more useful than specifically relevant example problems for understanding the structure of novel problems because these general representations are not contaminated by specific content (Holyoak, 1985). Also Z. Chen and Klahr (1999) showed that training students in how to conduct experiments that allow for causal inferences led to an increase in the knowledge acquired, even though it was gathered in a specific context (i.e., science education). This knowledge was successfully transferred to different tasks. Specifically, students in the experimental group performed better on tasks that were comparable to the original one but also in generalizing the knowledge gained across various tasks (Z. Chen & Klahr, 1999).

In line with Z. Chen and Klahr (1999), the results of our study also support the concept of generally important and transferable CPS processes. Changes in students' CPS performance may very well be reflected by corresponding increases in MicroDYN scores, independent of whether they are induced by specific training methods such as guided discovery or by school in general. Therefore, we suggest thoroughly investigating the educational implications of using MicroDYN as a training tool for domain-unspecific knowledge acquisition and application skills. It is under this assumption that CPS is employed in PISA 2012 as a domain-general and complementary measure to domain-specific concepts (OECD, 2010).

However, even though today's students need to be prepared to meet different challenges than those of 30 years ago, and even though the concept of life-long learning, which extends the educational span to a lifetime, has become increasingly popular (M. C. Smith & Reio, 2006), one should not count one's chickens before they hatch. Said otherwise, it may be premature to consider specific training issues. Further, a deeper understanding of CPS and its relation to *g* seems to be needed in light of the scarce empirical evidence. With the present study, we want to empirically and conceptually contribute to this new debate, and we conclude by emphasizing the great potential that CPS has as an educationally relevant construct. Just as Alfred Binet and Théodore Simon (1904) saw the relevance of general mental ability for academic achievement and laid the foundation of modern intelligence research, Gestalt psychologists such as Karl Duncker (1945) were well aware of the implications and importance of problem solving in education. However, it is only in light of current developments that the issue of how to make students good problem solvers is finally receiving the attention it deserves within psychology.

References

- Alsbaugh, J. W., & Harting, R. D. (1995). Transition effects of school grade-level organization on student achievement. *Journal of Research & Development in Education*, 28(3), 145–149.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: AERA.
- Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*, 22(1), 121–130. doi:10.1037/a0017767
- Binet, A., & Simon, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anomaux [New methods for assessing the intellectual level of anormal individuals]. *L'Année Psychologique*, 11, 191–244. doi:10.3406/psy.1904.3675
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale, NJ: Erlbaum.
- Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, 13(2), 287–321. doi:10.1207/s15328007sem1302_7
- Byrnes, J. P. (2001). *Minds, brains, and education*. New York, NY: Guilford Press.
- Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). Amsterdam, the Netherlands: Pergamon.
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95, 1005–1018. doi:10.1037/a0013193
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098–1120. doi:10.1111/1467-8624.00081
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. doi:10.1207/S15328007SEM0902_5
- Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (2011). Beyond IQ: A latent state trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence*, 39(5), 323–334. doi:10.1016/j.intell.2011.06.004
- Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement. The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, 19(2), 294–304. doi:10.1037/0893-3200.19.2.294
- Deary, I. J., Whalley, L. J., & Crawford, J. R. (2004). An “instantaneous” estimate of a lifetime's cognitive change. *Intelligence*, 32, 113–119. doi:10.1016/j.intell.2003.06.001
- Diaz, L., & Heining-Boynton, A. L. (1995). Multiple intelligences, multiculturalism, and the teaching of culture. *International Journal of Educational Research*, 23, 607–617. doi:10.1016/0883-0355(96)80440-X
- Dörner, D. (1986). Diagnostik der operativen Intelligenz [Assessment of operative intelligence]. *Diagnostica*, 32(4), 290–308.
- Dörner, D. (1990). The logic of failure. In D. E. Broadbent, J. T. Reason, & A. D. Baddeley (Eds.), *Human factors in hazardous situations* (pp. 15–36). New York, NY: Oxford University Press.
- Dörner, D., & Kreuzig, W. (1983). Problemlösefähigkeit und Intelligenz [Problem solving ability and intelligence]. *Psychologische Rundschau*, 34, 185–192.
- Duncker, K. (1945). On problem solving. *Psychological Monographs*, 58(5, Whole No. 270).
- Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving*, 4(1), 19–42.
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, 7(1), 69–89. doi:10.1080/13546780042000046
- Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, 11, 133–142. doi:10.1007/s10339-009-0345-0
- Funke, J., & Frensch, P. A. (2007). Complex problem solving: The Euro-

- pean perspective—10 years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25–47). New York, NY: Erlbaum.
- Gardner, P. H., & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology*, 9(7), S55–S79. doi:10.1002/acp.2350090706
- Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York, NY: Bantam Books.
- Greeno, J. G., & Simon, H. A. (1988). Problem solving and reasoning. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (2nd ed., Vol. 2, pp. 589–672). Hillsdale, NJ: Erlbaum.
- Greiff, S. (2012). *Individualdiagnostik der Problemlösefähigkeit* [Diagnostics of problem solving ability on an individual level]. Münster, Germany: Waxmann.
- Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new measurement perspective. *Applied Psychological Measurement*, 36(3), 189–213. doi:10.1177/0146621612439620
- Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, 8, 179–203. doi:10.1016/0160-2896(84)90008-4
- Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 59–87). New York, NY: Academic Press.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Jensen, A. R. (1998). The g factor and the design of education. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, instruction, and assessment* (pp. 111–131). Mahwah, NJ: Erlbaum.
- Kihlstrom, J. F., & Cantor, N. (2011). Social intelligence. In R. J. Sternberg & S. C. Barry (Eds.), *The Cambridge handbook of intelligence* (pp. 564–581). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511977244.029
- Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, 33(4), 347–368. doi:10.1016/j.intell.2005.03.002
- Kyllonen, P. C. (2009). New constructs, methods, and directions for computer-based assessment. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment—Lessons learned from large-scale surveys and implications for testing* (pp. 151–156). Luxembourg City, Luxembourg: Office for Official Publications of the European Communities.
- Lievens, F., & Chan, D. (2010). Practical intelligence, emotional intelligence, and social intelligence. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 339–359). New York, NY: Routledge.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9(2), 151–173. doi:10.1207/S15328007SEM0902_1
- Lynch, M., & Macbeth, D. (1998). Demonstrating physics lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 269–297). Mahwah, NJ: Erlbaum.
- Mansfield, R. S., Busse, T. V., & Krepelka, E. J. (1978). The effectiveness of creativity training. *Review of Educational Research*, 48, 517–536.
- Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.
- Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River, NJ: Prentice Hall.
- Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 287–303). Mahwah, NJ: Erlbaum.
- McGrew, K. S. (2009). CHC theory and the Human Cognitive Abilities Project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1–10. doi:10.1016/j.intell.2008.08.004
- Molnár, G., & Csapó, B. (2007, August 28–September 1). *Constructing complex problem solving competency scales by IRT models using data of different age groups*. Abstract submitted at the 12th Biennial EARLI Conference, Budapest, Hungary.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.
- Myrberg, E., & Rosen, E. (2008). A path model with mediating factors of parents' education on students' reading achievement in seven countries. *Educational Research and Evaluation*, 14(6), 507–520. doi:10.1080/13803610802576742
- Nagy, J. (2008). Renewing elementary education. In K. Fazekas, J. Köllö, & J. Varga (Eds.), *Green book for renewal of public education in Hungary* (pp. 61–80). Budapest, Hungary: Ecostat.
- Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51, 77–101. doi:10.1037/0003-066X.51.2.77
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.
- Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 321–349). Cambridge, England: Cambridge University Press.
- Novick, L. R., Hurley, S. M., & Francis, M. (1999). Evidence for abstract, schematic knowledge of three spatial diagram representations. *Memory & Cognition*, 27, 288–308. doi:10.3758/BF03211413
- Organisation for Economic Co-operation and Development. (2004). *Problem solving for tomorrow's world. First measures of cross-curricular competencies from PISA 2003*. Paris, France: OECD.
- Organisation for Economic Co-operation and Development. (2009). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics, and science* (Vol. 1). Paris, France: OECD.
- Organisation for Economic Co-operation and Development. (2010). *PISA 2012 problem solving framework* [draft for field trial]. Paris, France: OECD.
- Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [On the relation between test intelligence and problem solving success]. *Zeitschrift für Psychologie*, 189, 79–100.
- Raven, J. C. (1962). *Advanced progressive matrices, Set II*. London, England: H. K. Lewis.
- Raven, J. (2000). Psychometrics, cognitive ability, and occupational performance. *Review of Psychology*, 7, 51–74.
- Ree, H. M., & Carretta, T. R. (2002). g2K. *Human Performance*, 15, 3–24.
- Reeve, C. L., & Hakel, M. D. (2002). Asking the right questions about g. *Human Performance*, 15, 47–74. doi: 10.1080/08959285.2002.9668083
- Rigas, G., Carling, E., & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence*, 30, 463–480. doi: 10.1016/S0160-2896(02)00121-6
- Rindermann, H., Flores-Mendoza, C., & Mansur-Alves, M. (2010). Reciprocal effects between fluid and crystallized intelligence and their dependence on parents' socioeconomic status and education. *Learning and Individual Differences*, 20(5), 544–548. doi:10.1016/j.lindif.2010.07.002
- Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979–2002. *Journal of School Psychology*, 40(6), 451–475. doi:10.1016/S0022-4405(02)00124-3
- Salthouse, T. A., & Davis, H. P. (2006). Organization of cognitive abilities and neuropsychological variables across the lifespan. *Developmental Review*, 26, 31–54. doi:10.1016/j.dr.2005.09.001
- Smith, M. C., & Reio, T. G. (2006). Adult development, schooling, and the transition to work. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 115–138). Mahwah, NJ: Erlbaum.

- Smith, S. S. (2006). Examining the long-term impact of achievement loss during the transition to high school. *Journal of Secondary Gifted Education, 17*(4), 211–221.
- Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., . . . Latour, T. (2012). The Genetics Lab. Acceptance and psychometric characteristics of a computer-based microworld to assess complex problem solving. *Psychological Test and Assessment Modeling, 54*(1), 54–72.
- Spearman, C. (1927). *The abilities of man. Their nature and measurement*. New York, NY: Macmillan.
- Sternberg, R. J. (2000). The holey grail of general intelligence. *Science, 289*, 399–401. doi:10.1126/science.289.5478.399
- Sternberg, R. J. (2009). Toward a triachic theory of human intelligence. In R. J. Sternberg, J. C. Kaufman, & E. L. Grigorenko (Eds.), *The essential Sternberg: Essays on intelligence, psychology, and education* (pp. 38–70). New York, NY: Springer. (Original work published 1984)
- Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem-solving ability. *Educational Measurement: Issues and Practice, 14*(3), 29–35. doi:10.1111/j.1745-3992.1995.tb00865.x
- Taub, G. E., & McGrew, K. S. (2004). A confirmatory factor analysis of Cattell-Horn-Carroll theory and cross-age invariance of the Woodcock-Johnson Tests of Cognitive Abilities III. *School Psychology Quarterly, 19*(1), 72–87. doi:10.1521/scpq.19.1.72.29409
- Ullman, J. B. (2007). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (pp. 676–780). Boston, MA: Pearson.
- Wechsler, D. (2008). *Wechsler Adult Intelligence Scale 4th edition*. San Antonio, TX: Pearson Assessment.
- Weiß, R. H. (2006). *Grundintelligenztest Skala 2—Revision (CFT 20-R)* [Culture Fair Intelligence Test 20-R—Scale 2]. Göttingen, Germany: Hogrefe.
- Wigfield, A., Byrnes, J. P., & Eccles, J. S. (2006). Development during early and middle adolescence. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 87–113). Mahwah, NJ: Erlbaum.
- Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence, 40*, 1–14. doi:10.1016/j.intell.2011.11.003

Appendix

MicroDYN Item Characteristics and Linear Structural Equations

Item and linear structural equations	System size	Effects
Item 1 $X_{t+1} = 1 \cdot X_t + 0 \cdot A_t + 2 \cdot B_t$ $Y_{t+1} = 1 \cdot Y_t + 0 \cdot A_t + 2 \cdot B_t$	2×2 system	Only direct
Item 2 $X_{t+1} = 1 \cdot X_t + 2 \cdot A_t + 2 \cdot B_t + 0 \cdot C_t$ $Y_{t+1} = 1 \cdot Y_t + 0 \cdot A_t + 0 \cdot B_t + 2 \cdot C_t$	2×3 system	Only direct
Item 3 $X_{t+1} = 1 \cdot X_t + 0 \cdot A_t + 2 \cdot B_t + 0 \cdot C_t$ $Y_{t+1} = 1 \cdot Y_t + 2 \cdot A_t + 0 \cdot B_t + 0 \cdot C_t$ $Z_{t+1} = 1 \cdot Z_t + 0 \cdot A_t + 0 \cdot B_t + 2 \cdot C_t$	3×3 system	Only direct
Item 4 $X_{t+1} = 1 \cdot X_t + 2 \cdot A_t + 0 \cdot B_t + 0 \cdot C_t$ $Y_{t+1} = 1 \cdot Y_t + 0 \cdot A_t + 2 \cdot B_t + 2 \cdot C_t$ $Z_{t+1} = 1 \cdot Z_t + 0 \cdot A_t + 0 \cdot B_t + 2 \cdot C_t$	3×3 system	Only direct
Item 5 $X_{t+1} = 1 \cdot X_t + 2 \cdot A_t + 2 \cdot B_t + 0 \cdot C_t$ $Y_{t+1} = 1 \cdot Y_t + 0 \cdot A_t + 2 \cdot B_t + 0 \cdot C_t$ $Z_{t+1} = 1 \cdot Z_t + 0 \cdot A_t + 0 \cdot B_t + 2 \cdot C_t$	3×3 system	Only direct
Item 6 $X_{t+1} = 1.33 \cdot X_t + 2 \cdot A_t + 0 \cdot B_t + 0 \cdot C_t$ $Y_{t+1} = 1 \cdot Y_t + 0 \cdot A_t + 0 \cdot B_t + 2 \cdot C_t$	2×3 system	Direct and indirect
Item 7 $X_{t+1} = 1 \cdot X_t + 2 \cdot A_t + 0 \cdot B_t + 0 \cdot C_t$ $Y_{t+1} = 1 \cdot Y_t + 2 \cdot A_t + 0 \cdot B_t + 0 \cdot C_t$ $Z_{t+1} = 1.33 \cdot Z_t + 0 \cdot A_t + 0 \cdot B_t + 2 \cdot C_t$	3×3 system	Direct and indirect

Note. The seven items in this study were varied mainly on two system attributes proven to be most influential on item difficulty (see Greiff, 2012): the number of effects between variables and the quality of effects (i.e., effects of input and output variables). X_t , Y_t , and Z_t denote the values of the output variables, and A_t , B_t , and C_t the values of the input variables, during the present trial, whereas X_{t+1} , Y_{t+1} , and Z_{t+1} denote the values of the output variables in the subsequent trial.

Received December 16, 2011
Revision received November 5, 2012
Accepted December 26, 2012 ■

A Meta-Analysis of the Efficacy of Teaching Mathematics With Concrete Manipulatives

Kira J. Carbonneau, Scott C. Marley, and James P. Selig
University of New Mexico

The use of manipulatives to teach mathematics is often prescribed as an efficacious teaching strategy. To examine the empirical evidence regarding the use of manipulatives during mathematics instruction, we conducted a systematic search of the literature. This search identified 55 studies that compared instruction with manipulatives to a control condition where math instruction was provided with only abstract math symbols. The sample of studies included students from kindergarten to college level ($N = 7,237$). Statistically significant results were identified with small to moderate effect sizes, as measured by Cohen's d , in favor of the use of manipulatives when compared with instruction that only used abstract math symbols. However, the relationship between teaching mathematics with concrete manipulatives and student learning was moderated by both instructional and methodological characteristics of the studies. Additionally, separate analyses conducted for specific learning outcomes of retention ($k = 53$, $N = 7,140$), problem solving ($k = 9$, $N = 477$), transfer ($k = 13$, $N = 3,453$), and justification ($k = 2$, $N = 109$) revealed moderate to large effects on retention and small effects on problem solving, transfer, and justification in favor of using manipulatives over abstract math symbols.

Keywords: mathematics, manipulatives, concrete objects, activity-based learning, hands-on learning

Results from the 2011 National Assessment of Educational Progress (National Center for Education Statistics, 2011) indicate 60% of fourth-grade and 57% of eighth-grade United States students failed to meet standards of proficiency in mathematics. Furthermore, with only 10% of fourth graders and 6% of eighth graders meeting international standards of advanced proficiency, U.S. students rank below their same-age peers from eight countries (National Center for Education Statistics, 2008). These results, and comparable findings from prior years, have provided President Obama motivation for a recent executive branch initiative known as Educate to Innovate (The White House, Office of Press Secretary, 2009). This initiative was developed to target student achievement within science, technology, engineering, and math education with a focus on increasing domain-specific critical reasoning skills. If the goal of Educate to Innovate is to help students reach high levels of mathematics achievement, efficacious instructional strategies need to be identified.

Therefore, a careful examination of contemporary instructional strategies is necessary to identify strategies that improve mathematics achievement.

Instructional strategies that use manipulatives are often suggested as effective approaches to improve student mathematics achievement (Gürbüz, 2010; Sherman & Bisanz, 2009). Math manipulative-based instructional techniques are approaches that include opportunities for students to physically interact with objects to learn target information (Carbonneau & Marley, 2012). As examples, at the elementary level, teachers use play money to help students learn basic arithmetic functions, and at the high school level, teachers use plastic algebra tiles to teach concepts associated with division and multiplication within an equation. The National Council of Teachers of Mathematics (NCTM, 2000) has recommended that students be provided access to manipulatives in order to develop mathematical understanding. In addition, teacher education textbooks often contain sections suggesting that teachers use manipulatives during mathematics instruction (e.g., Billstein, Libeskind, & Lott, 2009; Copley, 2000). In the cases of national organizations and textbooks, when an instructional strategy is prescribed to a professional audience, an underlying assumption is that sound scientific evidence supports the recommendation. However, evidence supporting the efficacy of concrete math manipulatives is inconsistent. Specifically, the efficacy of manipulatives in mathematics instruction has not been uniformly observed with various populations, math topics, and cognitive outcomes. Variability of the effectiveness within instructional strategies of this nature may result in the misapplication of the instructional technique.

The instructional strategies literature is not definitive regarding the efficacy of concrete manipulatives. Studies have found that using manipulatives in math instruction, when compared with instruction that did not use manipulatives, may benefit student

This article was published Online First December 17, 2012.

Kira J. Carbonneau, Scott C. Marley, and James P. Selig, Department of Individual, Family and Community Education, Educational Psychology Program, College of Education, University of New Mexico.

Funding for this research was provided by the University of New Mexico's College of Education. Preliminary results were presented at the American Educational Research Association 2012 Annual Conference, Vancouver, Canada. We thank Carolyn J. Hushman for her assistance with coding studies.

Correspondence concerning this article should be addressed to Kira J. Carbonneau or Scott C. Marley, Department of Individual, Family and Community Education, Educational Psychology Program, College of Education, University of New Mexico, 117 Simpson Hall, MSC 05 3040, Albuquerque, NM 87131-1246. E-mail: kjcarbonneau@gmail.com or marley@unm.edu

learning (Gürbüz, 2010), result in comparable performance (Canobi, 2005; Dyer, 1996), or reduce student learning (Shoecraft, 1971). These contradictions may exist as a result of systematic factors. For instance, the level of instructional guidance, type of manipulative, age of learners, and other characteristics of a learning environment may impact the effectiveness of the intervention. Therefore, a systematic review of the math manipulatives literature is necessary to understand the variations in results observed between studies.

Sowell (1989) performed the first research synthesis of the manipulatives literature with a meta-analysis of the use of manipulatives applied to mathematics learning. Sowell's results suggested that relative to studies that did not use manipulatives, small-sized statistical differences in favor of the use of manipulatives on measures of recall were present when instruction was implemented over a school year. One limitation of Sowell's synthesis is that it did not examine whether instructional characteristics moderate the effectiveness of math manipulatives in terms of student learning. In addition to this limitation, there has been a considerable expansion of the math manipulatives literature base since Sowell's 1989 study. Both of these circumstances justify another systematic review of the math manipulatives literature. Therefore, the purposes of the present meta-analysis were to determine the average effect of math manipulatives and to identify potential moderators of the effectiveness of manipulatives. Instructional moderators were identified based on theoretical explanations for the efficacy of manipulatives, whereas methodological characteristics were examined to evaluate the credibility of the literature (Marley & Levin, 2011).

Moderators of the Efficacy of Manipulatives: Instructional Characteristics

Potential instructional moderators of the efficacy of teaching with manipulatives can be derived from contemporary human development and cognitive theories (McNeil & Jarvin, 2007). According to these theoretical explanations, concrete manipulatives facilitate learning by (a) supporting the development of abstract reasoning (Bruner, 1964; Montessori, 1964; Piaget, 1962), (b) stimulating learners' real-world knowledge (Baranes, Perry, & Stigler, 1989; Rittle-Johnson & Koedinger, 2005), (c) providing the learner with an opportunity to enact the concept for improved encoding (Kormi-Nouri, Nyberg, & Nilsson, 1994), and (d) affording opportunities for learners to discover mathematical concepts through learner-driven exploration (Bruner, 1961; Papert, 1980; Piaget & Coltman, 1974). Each of these theoretical explanations provides instructional characteristics that may reduce or increase the effectiveness of math manipulatives. The following sections describe each theoretical explanation and theory-relevant instructional factors.

Development of Abstract Reasoning

According to developmental theorists (Bruner, 1964; Montessori, 1964; Piaget, 1962), young children are expected to obtain cognitive benefits from exploring mathematical concepts with manipulatives. Empirical research examining concrete manipulatives are commonly situated through Piagetian developmental statuses with a focus on children at ages associated with concrete

operations (Fennema, 1972; Fujimura, 2001; Garcia, 2004). In addition, theoretical perspectives indicate that children in early childhood (age 7 and younger) should benefit from exploring mathematical concepts with manipulatives (Montessori, 1964). Children within both age groups are expected to derive greater cognitive benefits from manipulatives relative to older children (Fennema, 1972; Resnick & Omanson, 1987). The reason for this expectation is that younger children are assumed to have a greater dependency on physically interacting with their environment to construct meaning (Bruner, 1964; Piaget & Coltman, 1974). Through these physical interactions with the environment, young children are expected to gain proficiency with higher level representations in a predictable sequence. This sequence predicts that the ability for children to capitalize from visual representations should precede symbolic representations.

Inherent in these theoretical perspectives is the prediction that the developmental status of students should moderate the efficacy of teaching math topics with concrete manipulatives. It is expected that older students who have developed the ability to reason abstractly can benefit from instruction that consists exclusively of symbolic representations. Younger children, however, are predicted to experience more difficulty when provided instruction that solely consists of symbolic representation. Therefore, the assumed cognitive benefits of manipulating concrete objects to represent mathematical concepts should be greater for younger children who are still developing proficiency with higher level representations.

Stimulating Real-World Knowledge

The use of manipulatives in mathematics instruction has been cited as a strategy to allow students to draw on their practical knowledge (Burns, 1996). This line of reasoning suggests that concrete objects that resemble everyday items should assist students in making connections between abstract mathematical concepts and the real world (Brown, McNeil, & Glenberg, 2009). Support for this argument is provided by evidence indicating that when prior knowledge of a concept is partial, or absent, providing known concrete objects may help learners construct context-relevant schemas (e.g., Tindall-Ford & Sweller, 2006). However, results from empirical research examining the connection between student learning and the type of manipulatives used during instruction has been counterintuitive and could potentially account for some of the inconsistencies within the manipulation-based literature. Research examining different types of manipulatives tends to focus on the perceptual richness of concrete objects and how details of an object may hinder or aid learning.

Research examining the perceptual richness of a manipulative has primarily focused on realism or visual details of manipulatives. These examinations compare realistic manipulatives (e.g., manipulatives that look like pizza or money) that are perceptually rich (McNeil, Uttal, Jarvin, & Sternberg, 2009) to manipulatives that are nondescript or bland in nature (e.g., manipulatives that represent geometric shapes or place value). Results from these studies suggest that perceptually rich manipulatives may hinder learning of targeted mathematics concepts and/or performance solving mathematics problems (Kaminski, Sloutsky, & Heckler, 2009; McNeil et al., 2009). Explanations for why learning is inhibited by perceptually rich manipulatives have focused on generalizing learning to other contexts (Martin, 2009), surface information that

is irrelevant to the target concept (Kaminski et al., 2009; McNeil et al., 2009), and children not recognizing that concrete objects can be representative of an actual object and abstract mathematics concepts (Uttal, O'Doherty, Newland, Hand, & DeLoache, 2009).

Enactment Effects

Instructional strategies that use manipulatives may be effective because of physical enactment. In other words, the encoding and subsequent retrieval of target information may occur via nonverbal coding or a motoric channel. A well-developed literature exists examining what are known as self-performed tasks (SPTs). SPTs are tasks that participants physically perform during a learning activity. Often examined in paired-associate or list-learning contexts, SPTs have been found to result in robust encoding that enhances subsequent retrieval of target information (Engelkamp, Zimmer, Mohr, & Sellen, 1994; Kormi-Nouri, Nyberg, & Nilsson, 1994).

Dual coding theory offers an explanation for the memory benefits of SPTs. According to this theory, verbal and nonverbal representations are stored in separate but connected stores in long-term memory (Paivio, 1986). Consequently, it is proposed that activation of one form of representation leads to the activation of the other, resulting in the improved retrieval of target information (Clark & Paivio, 1987). In an instructional context, a child studying math facts using manipulatives to represent quantities is learning the target concept with both forms of representation present. Later, when asked to remember the target information, the child would have access to a verbal code consisting of target math facts and a nonverbal code consisting of interactions with the manipulatives. The successful retrieval of one form of representation is expected to activate the other, which in turn should result in greater performance on learning outcomes (for relevant discussion, see Marley & Levin, 2006).

The process of enactment has been demonstrated as an efficacious learning strategy within the content area of reading comprehension. Manipulating objects as directed by a narrative has been found to improve memory for spatial relationships and story events (Biazak, Marley, & Levin, 2010). Likewise, the ability to benefit from imagery instruction has been associated with the enactment of manipulatives. For example, Glenberg, Gutierrez, Levin, Japuntich, and Kaschak (2004) found that interacting with text-relevant objects during reading instruction increased comprehension of stories when participants were subsequently asked to imagine manipulating the text-relevant objects. This finding has been replicated in studies with children from samples representing diverse populations (Marley, Szabo, Levin, & Glenberg, 2011; Marley & Szabo, 2010).

A potential problem with physical enactment has been identified by several authors (Martin, 2009; Sarama & Clements, 2009). The simple act of moving manipulatives is likely not sufficient for promoting learning. Without explicit instruction, children may not move objects in a manner that appropriately represents the mathematics concept being taught. In other words, the instructional guidance provided is expected to influence the efficacy of manipulatives and the process of engaging in SPTs.

Learner-Driven Exploration

Many have suggested that providing learners with opportunities to discover mathematical concepts through unstructured learner-driven exploration will result in robust learning outcomes (Bruner, 1961; Piaget & Colman, 1974). These theorists propose that learners are better able to construct meaningful knowledge when given opportunities to discover concepts (Lefrançois, 1997). Empirical research has provided evidence contradicting this notion with results indicating that providing learners with instructional guidance on topics rather than allowing them to work within a purely unstructured context results in higher levels of student performance (Mayer, 2004).

Instructional guidance offered to learners can be defined as the amount of instructional support provided during the learning process and falls on a continuum of student- versus teacher-controlled learning (Kirschner, Sweller, & Clark, 2006; Mayer, 2004). On one end of this continuum are student-controlled strategies that allow learners to use manipulatives in an unstructured or less structured environment (i.e., a low guidance or discovery learning environment). Students in low guidance scenarios, often identified as math explorations, are given manipulatives with little or no instruction on how to manipulate the objects to represent mathematical concepts under study (Hinzman, 1997; Kuhfittig, 1974; LeBlanc, 1968). On the other end of the continuum are teacher-controlled strategies in which students interact with manipulatives as instructed by a teacher (i.e., direct instruction).

A recent synthesis of the instructional guidance literature indicates the provision of instructional guidance results in greater performance on learning outcomes relative to pure discovery (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011). Reading and listening strategy research further supports the importance of instructional guidance when using concrete manipulatives (Glenberg, Brown, & Levin, 2007; Marley, Levin, & Glenberg, 2007, 2010; Marley et al., 2011). However, Martin (2009) warned that too much instructional guidance with concrete manipulatives can impede learning by confining students to interpretations that do not transfer to novel circumstances. If this so, it is expected that the provision of high instructional guidance with manipulatives will result in lower performance on outcomes related to transfer of learning.

Moderators of the Efficacy of Manipulatives: Methodological Characteristics

Methodological aspects of a study affect the credibility of the claims that can be made regarding the causal relationship between an instructional strategy and beneficial learning outcomes (for relevant discussions, see Marley & Levin, 2011; Shadish, Cook, & Campbell, 2002). The robustness of claims derived from a literature can be assessed by examining the prevalence of studies that have characteristics linked with limitations in internal (e.g., pre- and postdesigns) and external (e.g., researcher-implemented treatments) validities. Whether these factors are associated with effect size is of interest, because educational recommendations should be based upon a body of scientifically credible evidence. For the present meta-analytic review, potential moderators of interest were whether design (pre- and postdesign, quasi-experiment, true experiment), type of test (standardized test, researcher-created tests), assumption of statistical independence (accounted for in analysis,

unaccounted for in analysis), implementer (researcher, teacher), and peer review status (published, unpublished) were associated with effect size. Although peer review status is not a methodological characteristic, it is included because acceptance in a peer-reviewed publication may serve as a proxy for a study's methodological rigor.

Present Study

The purpose of the present study was to ascertain the effectiveness of using manipulatives to teach mathematics when compared with teaching mathematics with only abstract math symbols. Moderators were examined to determine whether efficacy of this strategy differed by instructional and methodological characteristics. In summary, we sought to answer the following research questions:

1. What is the average effect of using concrete manipulatives in mathematics instruction?
2. Does the relationship between using concrete manipulatives and student learning vary by learning outcome?
3. Do instructional characteristics of studies moderate the relationship between using concrete manipulatives and learning outcomes?
4. Do methodological characteristics of studies moderate the relationship between using concrete manipulatives and learning outcomes?

In addition to addressing the primary research questions, we evaluated the overall quality of the literature by describing the prevalence of specific design characteristics. For a literature base to be considered robust in terms of validity of causal inferences and generalizability, it should primarily consist of studies that begin with random assignment and are representative of instructional contexts.

Method

Literature Search

An exhaustive search for studies on manipulatives and mathematics was performed between August 2010 and March 2011 with relevant keywords and their combinations (e.g., *mathematics, manipulatives, concrete objects, activity-based learning, hands-on learning*) with six major databases in the social sciences (Education Resources Information Center, Education Research Complete, PsycARTICLES, PsycINFO, JSTOR, and ProQuest Digital Dissertations). This search resulted in the identification of 94 articles. From these articles, ancestral and secondary citations from the studies' references were collected and examined for relevance to the present study. The search of ancestral and secondary citations resulted in 102 additional references ($k = 196$). Of the 196 studies, projects that did not report empirical findings were removed. This resulted in 101 studies to be reviewed for inclusion in the meta-analysis.

Criteria for Inclusion

Four conditions were established to restrict studies to those that empirically examined the efficacy of manipulatives in mathematics instruction. First, to be included, a study was required to compare an instructional technique that used manipulatives with a comparison group that taught math with only abstract math symbols. This comparison group was defined by the following attributes: (a) no manipulatives were present, (b) all students were taught the same math concept, and (c) no iconic representations (e.g., pictures of base-10 blocks or virtual manipulatives) were present. Of the 101 studies, 21 (20.7%) failed to meet these criteria. This inclusion criterion produces a very specific comparison in which conditions where students physically interacted with concrete objects were compared with conditions where students were solely taught mathematics concepts with abstract math symbols.

Second, to be included in the meta-analysis, the examined instructional treatments must have provided some form of instruction during which students were able to learn from the manipulatives. Studies that examined only the performance of students with manipulatives were excluded; four studies (3.9%) failed to meet this criterion. The third criterion is based on the definition of manipulatives; studies that required students to work with rulers, scales, or calculators were not included, as these were seen as tools rather than manipulatives (8.0%, $k = 9$). Lastly, studies had to provide sufficient quantitative information to estimate an effect size, which resulted in the elimination of 11 studies (10.8%). Additionally, two publications reported results from the same study; therefore only one was retained in the sample. The screening process resulted in a total of 55 studies upon which all meta-analytic procedures were conducted.¹ Studies meeting the inclusion criteria along with summative information on design and findings are presented in Table 1.

Study Coding

All 55 studies were coded with a standardized protocol. The protocol was developed iteratively as studies were accumulated to include the moderators of interest. After a final coding scheme was developed, studies were coded for the characteristics described below. Two raters independently coded an overlapping random sample of 18 studies (32%) to assure consistency in coding. Interrater agreement, as measured by Cohen's κ for the categorical variables, ranged from .82 to 1.0. Pearson's r for continuous variables ranged from .89 to 1.0.

Effect size. Cohen's d (1988), a measure of effect size, was calculated for each study in the meta-analysis. Cohen's d values are obtained by dividing the difference of the treatment means by the pooled standard deviation. This measure of effect size is commonly reported in studies examining the effect of a manipulated independent variable (e.g., manipulatives vs. control) on a continuous dependent variable (e.g., retention, problem solving, transfer, or justification). When study statistics were not directly

¹ Sowell's 1989 meta-analysis of the math manipulative research consisted of 60 studies, of which approximately 23 contrasted pictures versus a symbolic representation. The publication does not include a list of reviewed studies. Sowell kindly responded to a request for a list of reviewed studies. She no longer had access to the list.

Table 1
Summary of Study Characteristics

Study	<i>N</i>	Days	Design ^a	Coded instructional characteristics ^b	Mean Cohen's <i>d</i>	Main findings
Aburime (2007)	185	50	QEX	FO, PRM, HG, GE, GI	0.01	Students in high school taught geometry with manipulatives performed the same on a measure of retention as those not taught with manipulatives.
Anderson (1957)	541	40	QEX	FO, LG, AR, GI,	0.08	Eighth-grade students taught algebra with manipulatives performed the same on a measure of retention as those taught with a textbook.
Aurich (1963)	90	180	QEX	PO, BM, HG, AR, GI	0.89	First graders taught arithmetic with Cuisenaire rods performed better on measures of reasoning and retention than those taught with a textbook.
Babb (1975)	76	25	QEX	CO, PRM, LG, AR, GI	-0.05	Second graders taught arithmetic with manipulatives performed the same on a measure of retention as those taught without manipulatives.
Battle (2007)	16	5	QEX	PO, PRM, LG, AR, GI	-0.93	First graders taught addition and subtraction with counters performed worse on a measure of retention than those taught without counters.
Bring (1972)	102	15	EX	CO, PRM, LG, AL, II	0.28	Fifth and sixth graders taught algebra with manipulatives performed better on a measure of retention than those students taught with a textbook.
Butler et al. (2003)	50	10	WS	CO, BM, HG, FR, GI	2.24	Mathematic disabled seventh graders taught fractions with manipulatives answered more items correctly on a postmeasure of retention.
Carmody (1970)	96	11	QEX	CO, PRM, HG, AR, GI	0.43	Sixth graders taught arithmetic with manipulatives performed better on a measure of transfer than students taught without manipulatives. Performance on a measure of retention was the same for both groups.
Cook (1967)	66	120	QEX	PO, LG, AR, GI	0.08	First-grade students taught arithmetic with manipulatives performed the same on a measure of problem solving than those taught with a textbook.
Cramer et al. (2002)	1,666	30	QEX	CO, PRM, HG, FR, GI	0.88	Fourth- and fifth-grade students taught fractions with manipulatives performed better on a measure of retention than those taught with a textbook.
Dawson (1955)	280	22	QEX	CO, PRM, LG, AR, GI	0.58	Fourth-grade students taught division with manipulatives performed better on a measure of retention than students who were taught without manipulatives.
Dyer (1996)	90	Unknown	QEX	FO, BM, LG, AL, GI	0.00	College-level students taught algebra with algebra tiles performed the same on a measure of retention than students taught with a textbook.
Egan (1990)	81	180	QEX	CO, BM, LG, AR, GI	-0.30	Second graders taught arithmetic with Cuisenaire rods performed worse on a measure of retention than those who were taught with a textbook.
Ekman (1967)	196	18	QEX	CO, PRM, HG, AR, GI	0.09	Third graders taught arithmetic with manipulatives performed the same on measures of retention and transfer as those taught with a textbook.

Table 1 (*continued*)

Study	<i>N</i>	Days	Design ^a	Coded instructional characteristics ^b	Mean Cohen's <i>d</i>	Main findings
Fennema (1972)	95	14	EX	CO, BM, HG, AR, GI	−0.65	Second graders taught arithmetic with manipulatives performed worse on a measure of transfer than students taught with a textbook.
Fujimura (2001)	76	I	EX	CO, PRM, LG, AL, II	0.73	Fourth graders taught proportions with manipulatives performed better on a measure of retention than those taught without manipulatives.
Garcia (2004)	64	20	QEX	CO, LG, AR, GI	−0.14	Third and fourth graders taught arithmetic with manipulatives performed the same on a measure of retention as those taught without manipulatives.
Getgood (2000)	287	10	QEX	CO, BM, HG, AL, GI	0.07	Sixth graders taught algebra with factor blocks performed better on a measure of immediate retention than those taught with a textbook. Performance on a second measure of retention was the same for both groups.
Goins (2001)	30	Unknown	QEX	FO, BM, HG, AL, GI	1.21	Ninth-grade students who were taught algebra with algebra tiles performed better on a postassessment than students who did not have access to the tiles.
Gürbüz (2010)	80	7	EX	CO, BM, HG, FR, GI	3.11	Seventh-grade students taught fractions with manipulatives performed better at word problems pertaining to probability than students taught without manipulatives.
Hawkins (1982)	35	20	WS	CO, PRM, HG, FR, GI	1.60	Third-grade students taught fractions with manipulatives answered more items correctly on a postmeasure of retention.
Hiebert et al. (1991)	25	11	WS	CO, BM, HG, PV, GI	0.47	Fourth-grade students taught decimals with manipulatives answered more items correctly on a postmeasure of retention with manipulatives. This gain was not seen on a measure of retention performed without manipulatives.
Hinzman (1997)	34	45	QEX	FO, BM, LG, AL, GI	0.20	College-level students taught algebra with manipulatives performed the same on a measure of retention than students who were not provided manipulatives.
Johnson (1970)	64	20	EX	CO, PRM, LG, AL, II	0.99	Fourth- and fifth-grade students taught prealgebra with manipulatives performed better on measures of retention than students taught without manipulatives.
Jordan et al. (1999)	125	Unknown	QEX	CO, PRM, HG, FR, GI	0.78	Fourth-grade students taught fractions with manipulatives performed better on measures of retentions than students who were taught without manipulatives.
King (1976)	134	5.5	EX	CO, PRM, HG, FR, GI	−0.02	Fourth-grade students taught fractions with manipulatives performed the same on measures of retention and transfer as those students taught with a textbook.
Kuhfittig (1974)	40	2	EX	CO, LG, AR, GI	−0.02	Seventh-grade students taught arithmetic with manipulatives performed the same on measures of retention and transfer as those students who were taught without manipulatives.

(table continues)

Table 1 (continued)

Study	<i>N</i>	Days	Design ^a	Coded instructional characteristics ^b	Mean Cohen's <i>d</i>	Main findings
Lucas (1966)	104	50	QEX	PO, PRM, LG, AR, GI	0.24	First-grade students taught with attribute blocks performed better on a measure of problem solving than students who were not taught with manipulatives.
Lucow (1964)	254	30	QEX	CO, BM, HG, AR, GI	0.76	Third-grade students taught arithmetic with manipulatives scored higher on a measure of retention than students taught with a textbook.
McClung (1998)	47	45	QEX	FO, BM, LG, AL, GI	−0.70	Tenth- and 11th-grade students taught algebra with manipulatives performed worse on a measure of retention than students taught without manipulatives.
Miller (1964)	114	9	WS	CO, PRM, HG, FR, GI	3.90	Sixth-grade students taught fractions with manipulatives answered more items correctly on a postmeasure of retention.
Moody et al. (1971)	90	20	EX	CO, PRM, HG, AR, GI	−0.07	Third-grade students taught multiplication with manipulatives performed worse on measures of retention and transfer than students taught with a textbook.
Nasca (1966)	45	180	QEX	CO, BM, HG, AR, GI	0.43	Second-grade students taught arithmetic with rods performed the same on measures of retention and transfer as those students taught with a textbook.
Nichols (1972)	267	Unknown	WS	CO, PRM, LG, AR, GI	2.12	Third-grade students taught multiplication with manipulatives answered more items correctly on a postmeasure of retention.
Nickel (1971)	90	30	QEX	CO, PRM, HG, AR, GI	0.40	Fourth-grade students taught arithmetic with manipulatives performed the same as students taught without manipulatives.
Nishida (2007)	78	1	EX	PO, BM, HG, FR, GI	0.11	First-grade students taught fractions with manipulatives performed the same on a measure of retention as students taught without manipulatives.
Norman (1955)	24	10	QEX	CO, HG, AR, GI	1.48	Third-grade students taught division with manipulatives performed better than students taught with a textbook.
Olkun (2003)	93	2	QEX	CO, PRM, LG, GE, II	0.37	Fourth- and fifth-grade students taught geometry with manipulatives performed better on a measure of retention than students taught without manipulatives.
Paolini (1977)	26	15	WS	PO, LG, AR, GI	0.88	Kindergarten students taught arithmetic with manipulatives answered more items correctly on a postmeasure of retention.
Peterson et al. (1988)	24	9	EX	CO, PRM, HG, PV, GI	0.67	Learning disabled fourth graders taught place value with manipulatives performed better on measures of retention and transfer than students taught without manipulatives.
Prigge (1978)	146	10	QEX	CO, PRM, LG, GE, II	0.59	Third-grade students taught geometry with manipulatives performed better on measures of retention than students taught with a textbook.
Robinson (1978)	119	5	EX	CO, BM, HG, FR, GI	0.27	Fourth graders taught fractions taught fractions with Cuisenaire rods performed the same on a measure of retention as students taught without rods.

Table 1 (*continued*)

Study	<i>N</i>	Days	Design ^a	Coded instructional characteristics ^b	Mean Cohen's <i>d</i>	Main findings
Shoecraft (1971)	1096	10	EX	FO, LG, FR, GI	−0.04	Seventh- and ninth-grade students taught fractions with manipulatives performed the same on measures of retention and problem solving as students taught without manipulatives.
Slaughter (1980)	217	Unknown	QEX	CO, BM, LG, AR, GI	0.18	Third- and fifth-grade taught arithmetic with manipulatives performed the same on measures of retention as students taught with a textbook.
Smith & Montani (2008)	12	12	WS	CO, BM, HG, AR, GI	0.84	Third graders taught multiplication with manipulatives answered more questions correctly on a postmeasure or retention.
Threadgill-Sowder & Juilfs (1980)	147	3	EX	CO, PRM, LG, GE, GI	0.47	Seventh-grade students taught geometry with manipulatives performed better on a measure of retention than students taught without. Performance on a measure of transfer favored students who were taught without manipulatives.
Steen et al. (2006)	31	15	WS	PO, HG, GE, GI	1.70	First-grade students taught geometry with manipulatives answered more items correctly on a postmeasure of retention.
Steger (1977)	52	12	WS	PO, BM, HG, PV, GI	0.63	First-grade students taught place value with manipulatives answered more items correctly on a postmeasure of retention.
Suh & Moyer (2007)	36	5	WS	CO, MB, HG, FR, GI	2.76	Third-grade students taught fractions with manipulatives answered more items correctly on a postmeasure of retention.
Taylor (2001)	58	3	QEX	CO, PRM, LG, AL, GI	−0.96	Fifth-grade students taught probability with manipulatives performed worse on a measure of retention than students taught with a textbook. Performance on a transfer measure was the same for both groups.
Wallace (1974)	154	Unknown	QEX	CO, BM, HG, AR, GI	1.42	Fifth-grade students taught arithmetic with Cuisenaire rods performed better on a measure of retention than students taught with a textbook.
Weber (1970)	30	30	QEX	PO, PRM, HG, AR, GI	0.18	First-grade students taught arithmetic with manipulatives performed the same on a measure of retention as students taught with a textbook.
Witzel et al. (2003)	68	20	QEX	CO, BM, HG, FR, GI	0.68	Sixth- and seventh-grade students taught fractions with manipulatives performed better on a measure of retention than students taught without manipulatives.
Wood (1974)	40	4	WS	CO, PRM, LG, AR, GI	0.12	Second-grade students taught multiplication with manipulatives answered the same amount of questions correctly on a postmeasure of retention and transfer.
Yuan et al. (2010)	60	4	WS	FO, PRM, LG, GE, GI	0.72	Eighth-grade students taught geometry with manipulatives answered more questions correctly on a postassessment of problem solving.

^a EX = experiment; QEX = quasi-experiment; WS = within subjects; ^b PO = preoperational students; CO = concrete operational students; FO = formal operational students; PRM = perceptually rich manipulatives; BM = bland manipulatives; HG = high instructional guidance; LG = low instructional guidance; PV = place value; AR = arithmetic; GE = geometry; FR = fractions; AL = algebra; GI = group instruction; II = individual instruction.

reported, Cohen's d values were calculated with reported descriptive statistics or observed F or t statistics (Rosenthal, 1984). Given that studies with larger samples should have a more precise estimate of the effect of manipulatives, studies were weighted to allow large samples to have more influence. Weights for each study were calculated from the reciprocal of the computed variance for d (for details, see Lipsey & Wilson, 2001).

A number of studies measured several outcomes of interest. Studies with various outcomes allowed for the coding of multiple effect sizes. For example, the study by Aurich (1963) measured three of the four coded learning outcomes (retention, problem solving, and transfer). This afforded the calculation of three effect sizes. To avoid the potential for nonindependence of effect sizes, effect sizes from the same study were averaged (Rosenthal & Rubin, 1986) to extract one effect size for each study. Effect sizes were also disaggregated to examine differential effects of manipulatives across learning outcomes.

Instructional Moderators

Developmental status. Both age and grade level were coded from each study. In cases with more than one grade level or age present, the mean was recorded. Subsequently, the variable of age was grouped into three categories equivalent to Piaget's stages of development, with samples of students from ages 3–6 coded as preoperational, ages 7–11 coded as concrete operational, and age 12 and older coded as formal operational.

Perceptual richness. Manipulatives were coded as either perceptually rich or bland in nature. A perceptually rich manipulative is defined as an object that is either representative of a real object or the actual object. For example, toy pizzas were coded as a perceptually rich manipulatives (e.g., Ekman, 1967; Peterson, Mercer, & O'Shea, 1988). Bland manipulatives included objects that are nondescript such as plain rectangular blocks or tangrams (e.g., Dyer, 1996; Egan, 1990).

Level of instructional guidance. Support provided to students was coded to represent two levels of instructional guidance. Studies were coded as providing either low or high instructional guidance. For instance, in a study done by Hinzman (1997), students worked in groups without instructions on how to use the manipulatives provided to them. The only instruction provided to students was the objective of the lesson, to represent equations using different manipulatives such as colored disks and plastic cups. In contrast, high instructional guidance was provided within Getgood's (2000) study that explicitly taught students the concepts of greatest common factor and least common multiple with factor blocks.

Mathematical topic. The math topic of instruction was coded as a categorical variable with the following five categories: place value, arithmetic, geometry, fractions, and algebra. The category of arithmetic was used as a broad term encompassing math operations such as addition, subtraction, multiplication, and division. These categories were developed to be exhaustive of all topics present in the current body of literature.

Group versus individual instruction. Implementation of manipulatives strategies was coded as a dichotomous variable with studies being implemented at either the individual or the group level. Studies that were conducted in both small groups and whole class instruction were coded as group level. Bring (1972) presented

an example of a study implemented at the individual level. In the study students were removed from the classroom and asked to individually complete a series of tasks or worksheets. As examples of classroom-level implementation, Slaughter (1980) and Battle (2007) examined the use of manipulatives with students who were provided instruction in a whole group classroom setting.

Instructional time. The time of treatment implementation was coded in days and then broken into tertiles to represent treatment times that were short (less than or equal to 14 days), medium (15–45 days), or long (greater than or equal to 46 days) in length. Several studies failed to report the duration of treatment. These studies were not included in the instructional time moderator analyses.

Outcome measure. The dependent variables from each study were coded into the following four classifications: retention, problem solving, transfer, and justification. Retention was defined as an outcome that required students to solve basic facts (for example from present sample, see, e.g., Smith & Montani, 2008). Problem solving included tasks in which students were not explicitly instructed on how to complete the assessment (see, e.g., LeBlanc, 1968; Shoecraft, 1971). Studies were classified as having a transfer outcome when students were asked to extend their knowledge to a new situation; for example, extending learned concepts of addition to multiplication (see, e.g., Ekman, 1967; Moody, Abell, & Bausell, 1971). Justification outcomes included activities in which participants were asked to provide explanations for why they used a given method to solve a problem (see, e.g., King, 1976).

Methodological Moderators

Peer review status. The publication source of each study was examined as a proxy variable for the quality of the study. Peer review status was coded as a dichotomous variable (published, not published). Studies were identified as published when they were located within a peer-reviewed journal. Other publication types such as thesis and dissertation studies found in research indexes (e.g., ProQuest) were coded as unpublished.

Research design. The design of each study was coded as either within-subjects, quasi-experimental, or experimental design. Studies classified as using within-subjects designs were those with a single group completing pre- and postassessments. The category of quasi-experimental designs included studies that manipulated the independent variable, but did not begin with random assignment to conditions. Studies with designs coded as experimental used random assignment to allocate individuals to conditions.

Implementer. Studies were also coded to identify who delivered treatment to students: a researcher or a teacher. When it was clear the person conducting the research was also the teacher of the students in the study, as was often the case in thesis and dissertation studies, the study was coded as treatment being delivered by the teacher.

Test type. Outcome measures were categorized as being researcher created or standardized assessments. For example, several studies used established assessments such as the Woodcock–Johnson (Woodcock & Johnson, 1989; see, e.g., Smith & Montani, 2008). Other studies used researcher-created materials, which were designed by the researcher specifically for use within the study (e.g., Hinzman, 1997).

Assumption of independence. The statistical assumption of independence was coded for each study. A dichotomous variable was created that distinguished studies that accounted for the non-independence of observations that can occur when students in the sample are nested within classrooms from those studies that did not. Studies were coded by examining the degrees of freedom used in the analysis. For instance, Garcia (2004) implemented the intervention at the group level and used the classroom as the unit of analysis, which meets the assumption of independence. In contrast, Prigge (1978) implemented at the classroom level, but the degrees of freedom for the analysis indicated that the unit of analysis was the individual, which does not meet the statistical assumption of independence.

Analyses

The analysis plan included separate, but parallel, analyses for the aggregated and disaggregated data. Initially the effect sizes were examined in the aggregated data set. For the aggregated data, studies that reported multiple effect sizes were assigned a single effect size that was the average of the reported or calculated effect sizes (Rosenthal & Rubin, 1986). This procedure was followed to avoid the issue of nonindependence that can arise when multiple effects sizes are nested within a study, and to better address questions regarding the overall efficacy of the use of manipulatives on student learning. For the aggregated data, our procedure included the following steps. First, we calculated a weighted mean effect size across all studies. Next, we examined between-study variation in effect sizes using a Q statistic (Hedges, 1983). If statistically significant levels of between-study variation were found, we examined moderation of effect sizes based on both substantive and methodological features of the studies. All moderator variables were categorical.

A partitioning of variance approach (Hedges, 1982; Hedges & Olkin, 1985) was used to examine moderation. This approach uses a Q_{between} statistic to represent the between-group variability in effect sizes. This value can be referenced against a chi-square distribution with $k - 1$ degrees of freedom to test a null hypothesis of no difference in effect sizes across levels of the moderator. When differences were found, and there were more than two levels of the moderator variable, we conducted post hoc pairwise comparisons using an extension of the Scheffé procedure (see Hedges & Olkin, 1985) to maintain family-wise error rates at .05. As a last step in our analysis of the aggregated data, we computed a fail-safe N (Rosenthal, 1979) to assess the possible impact of studies with nonsignificant findings being overlooked in the analysis.

Examination of the disaggregated data was parallel to that of the analysis of the aggregated data. For these analyses, four sets of effect sizes were calculated according to the type of outcome used within the study. This approach allowed us to address more specific questions regarding the effect of manipulatives used for different outcomes while also avoiding any nonindependence among the effect sizes (no study reported multiple effect sizes within a single category of outcomes). For each of the four types of outcomes, we calculated an overall effect size and tested the level of between-study variation. When statistically significant levels of between-study variation were found, moderation analyses, as described above, were conducted.

Results

Coded characteristics of the 55 summarized studies are presented in Tables 2 and 3. Findings illustrate important differences among the studies examining the efficacy of math manipulatives. Of note are the following details: Fifty-five percent were published in a peer-reviewed scholarly journal; 46% of the studies were done after Sowell's 1989 meta-analysis; 56% of the studies were with third- and fourth-grade children; 76% of the studies were quasi-experimental or pre- and postdesigns; 75% failed to account for the statistical assumption of independence; and in 73% of the studies classroom teachers implemented the intervention.

Aggregated Data

The aggregated mean effect size of 0.37 was statistically significant ($p < .001$, 95% CI [0.30, 0.44]). Hedges's homogeneity test for effect sizes was also statistically significant, $Q(54) = 277.8$, $p < .001$, suggesting that between-study variation in effect sizes exceeded what would be expected by sampling error alone.

Moderator analysis. Table 4 summarizes the findings from the analysis of the moderator variables. The effect of mathematical topic was found to be statistically significant, $Q(4) = 29.8$, $p < .001$. Post hoc comparisons indicated that the mean effect size for fractions ($d = 0.69$) was statistically greater than that for arithmetic ($d = 0.27$) and algebra ($d = 0.21$). Instructional guidance was also a statistically significant moderator, $Q(1) = 6.3$, $p = .01$, with the effect size of studies with high instructional guidance ($d = 0.46$) greater than those with low guidance ($d = 0.29$). In addition, developmental status was statistically significant, $Q(2) = 11.7$, $p = .002$; samples consisting of children assumed to be concrete operational had a greater mean effect size ($d = 0.45$) than those with samples consisting of participants assumed to be in formal operations ($d = 0.16$). The effect sizes for studies of preoperational ($d = 0.33$) students was not significantly different from the effect sizes for studies of either concrete or formal operational students. Lastly, the instructional variable of time was significant, $Q(2) = 9.8$, $p = .008$. Post hoc comparisons indicated that the mean effect sizes for instruction provided for short lengths of time (≤ 14 days) and medium lengths of time (15–45 days) were not statistically significantly different ($d = 0.34$ and $d = 0.45$, respectively). However, both mean effect sizes were significantly greater than the mean effect for long lengths (≥ 46 days; $d = 0.14$).

Several moderators based on methodological characteristics of studies were also statistically significant. Test type, $Q(1) = 3.8$, $p = .05$, was statistically significant, with studies using standardized assessments having a higher mean effect size ($d = 0.49$) than researcher-created assessments of learning ($d = 0.33$). The statistical assumption of independence, $Q(1) = 5.6$, was also statistically significant ($p = .01$), with studies that met the statistical

Table 2
Descriptive Statistics for Continuous Variables

Variable	M	SD	Range	
			Minimum	Maximum
Age	9.8 years	2.4	5.5	17.0
Treatment time	25.0 days	42.7	1.0	180.0

Table 3
Descriptive Statistics for Categorical Variables

Variable	Category	%
Peer review status	Not published	55.3
	Published	44.6
Instructional guidance	Low	44.6
	High	55.3
Independence	Not met	75.0
	Met	25.0
Test type	Researcher created	73.3
	Standardized assessment	26.7
Implementer	Teacher	73.3
	Researcher	26.7
Research design	Within subjects	23.2
	Quasi-experimental	53.5
	Experimental	23.2
Mathematical topic	Arithmetic	42.8
	Place value	5.4
	Geometry	10.7
	Fractions	23.2
	Algebra	17.8
Perceptual richness	Yes	53.0
	No	47.0
Group vs. individual instruction	Individual	7.1
	Group	92.9
Outcome measures	Retention	94.6
	Transfer	24.5
	Justification	3.7
	Problem solving	16.0

assumption having a smaller mean effect size ($d = 0.19$) than studies that did not meet this assumption ($d = 0.41$). Additionally, study design was statistically significant, $Q(2) = 91.5, p < .001$; post hoc comparisons revealed that within-subjects studies had a higher mean effect size ($d = 1.22$) than quasi-experimental studies ($d = 0.28$) or studies using the experimental design ($d = 0.16$). No statistical difference between quasi-experimental and experimental designs was observed. Finally, peer-reviewed status, $Q(1) = 5.7, p = .01$, was statistically significant, with published studies having a greater mean effect size ($d = 0.46$) than unpublished studies ($d = 0.30$).

Disaggregated Data

Analysis of the disaggregated effect sizes by the learning outcomes revealed a mean effect size for retention of 0.59 (95% CI [0.52, 0.65]), whereas the mean effect size for problem solving was 0.46 (95% CI [0.23, 0.68], with both being statistically significant ($p < .001$). The mean effect size for transfer was 0.13 (95% CI [0.02, 0.23]), and the mean effect size for justification was 0.38 (95% CI [0.06, 0.70], both $p < .01$). For three outcomes, examination of between-study variance revealed variation in effect sizes exceeded what would be expected by sampling error: retention, $Q(52) = 719.2$; transfer, $Q(12) = 56.7$; and problem solving, $Q(8) = 56.3$, all $p < .001$. For justification the Q statistic was found to be nonsignificant, $Q(1) = 2.93, p = .23$. Moderator analyses were performed only for the three outcome measures with statistically significant variation in effect sizes.

Retention. Table 5 summarizes the findings for each of the coded moderator variables within the learning outcome of retention. As with the aggregated data, variation among effect sizes was

impacted by both methodological and instructional variables. For instructional variables of interest, patterns similar to those found for the aggregated data emerged. Level of instructional guidance, $Q(1) = 106.5, p < .001$, was statistically significant, with high instructional guidance having a greater mean effect size ($d = 0.90$) than low instructional guidance ($d = 0.19$). Math topic was a statistically significant moderator, $Q(4) = 44.5, p < .001$, with studies concerning fractions ($d = 0.93$) and algebra ($d = 0.84$) having statistically higher mean effect sizes than studies teaching arithmetic ($d = 0.39$). Developmental status was significant, $Q(2) = 106.8, p < .001$, with post hoc comparisons revealing significant differences between all pairings of studies using pre-operational ($d = -0.09$), concrete operational ($d = 0.81$), and formal operational ($d = 0.31$) samples. Perceptual richness was a significant moderator, $Q(1) = 36.4, p < .001$, with studies that used perceptually rich objects ($d = 0.28$) having a lower mean effect size than studies with bland or nondescript objects ($d = 0.77$). Instructional time was also a significant moderator, $Q(2) = 7.4, p = .02$. Post hoc comparisons revealed that the mean effect size for instruction provided for short lengths of time ($d = 0.59$) was significantly greater than studies coded as medium lengths of time ($d = 0.35$), but not statistically different from long lengths of time ($d = 0.49$). Additionally, the difference between medium and long lengths of instructional time was not significant.

For methodological characteristics, several moderators were statistically significant. Peer review status was significant with published studies having a greater mean effect size ($d = 0.97$) than those that were unpublished ($d = 0.30$), $Q(1) = 90.7, p < .001$. Research design was significant, $Q(2) = 183.1, p < .001$. Studies using within-subjects designs had greater mean effect size ($d = 1.69$) than those using either quasi-experimental ($d = 0.35$) or experimental ($d = 0.47$) designs. The difference between studies using quasi-experimental and experimental designs was nonsignificant. Treatment implementer moderated effect size, $Q(1) = 89.1, p < .001$, with researcher-implemented treatments ($d = 0.13$) having a smaller mean effect size than teacher-implemented treatments ($d = 0.82$).

Problem solving. Table 6 contains the findings for the coded variables for studies measuring problem solving. Level of instructional guidance had a significant effect, $Q(1) = 19.1, p < .001$. High instructional guidance ($d = 1.06$) studies had a greater mean effect size than low instructional guidance studies ($d = 0.04$). Math topic also was a significant moderator, with $Q(3) = 45.1, p < .001$. Studies examining manipulatives with fractions had a greater mean effect size ($d = 2.50$) than those examining arithmetic ($d = 0.02$), place value ($d = 0.48$), and geometry ($d = 0.72$). The differences among the latter three were nonsignificant. Perceptual richness was significant, $Q(1) = 15.3, p < .001$, with studies that used perceptually rich objects ($d = -0.27$) having a lower mean effect size than studies with bland or nondescript objects ($d = 0.80$). Lastly, for instructional variables, time was a significant moderator, $Q(2) = 22.7, p < .001$, with instructional time coded as short ($d = 0.86$) and long ($d = 0.25$) in length having a greater mean effect size than studies coded as medium in length ($d = -0.62$). However, the difference between studies that were coded as short and long was not statistically significant.

Significant moderators related to research methods were present as well. Peer review status was significant, with published studies having a greater mean effect size ($d = 0.76$) than

Table 4
Variability of Effect Sizes Within Aggregated Data

Moderator	<i>k</i>	<i>N</i>	<i>d</i>	95% CI	<i>Q</i> _{between}
Methodology characteristics					
Peer review status					5.7, <i>p</i> = .01
Published	24	4,190	0.46	[0.36, 0.56]	
Not published	31	3,047	0.30	[0.22, 0.39]	
Design					91.5, <i>p</i> < .001
Within subject	12	748	1.22 _a	[1.03, 1.32]	
Quasi-experimental	30	5,173	0.28 _b	[0.20, 0.37]	
Experimental	13	1,346	0.16 _b	[0.02, 0.30]	
Implementer					1.6, <i>p</i> = .19
Researcher	15	1,285	0.29	[0.14, 0.43]	
Teacher	40	5,952	0.39	[0.32, 0.47]	
Test type					3.8, <i>p</i> = .05
Standardized	15	1,172	0.49	[0.35, 0.63]	
Researcher created	40	6,065	0.33	[0.26, 0.41]	
Independence					5.6, <i>p</i> = .01
Met	14	1,027	0.19	[0.03, 0.35]	
Not met	41	6,210	0.41	[0.33, 0.48]	
Instructional characteristics					
Instructional guidance					6.3, <i>p</i> = .01
High	30	4,275	0.46	[0.36, 0.56]	
Low	25	2,962	0.29	[0.20, 0.38]	
Mathematical topic					29.8, <i>p</i> < .001
Place value	3	101	0.58 _{ab}	[0.20, 0.96]	
Arithmetic	24	2,309	0.27 _a	[0.16, 0.38]	
Geometry	6	662	0.37 _a	[0.19, 0.56]	
Fractions	12	2,876	0.69 _b	[0.55, 0.84]	
Algebra	10	1,348	0.21 _a	[0.07, 0.34]	
Perceptual richness					0.21, <i>p</i> = .64
Yes	26	4,050	0.36	[0.27, 0.45]	
No	24	1,923	0.39	[0.28, 0.50]	
Group vs. individual					2.9, <i>p</i> = .08
Individual	4	388	0.58	[0.33, 0.83]	
Group	51	6,849	0.35	[0.29, 0.42]	
Development status					11.7, <i>p</i> = .002
Preoperational	10	1,256	0.33 _a	[0.16, 0.49]	
Concrete	38	5,657	0.45 _a	[0.37, 0.53]	
Formal	7	324	0.16 _b	[0.01, 0.31]	
Instructional time					9.4, <i>p</i> = .008
≤14 days	27	4,340	0.34 _a	[0.24, 0.45]	
15–45 days	16	1,353	0.45 _a	[0.34, 0.57]	
≥46 days	6	540	0.14 _b	[−0.01, 0.30]	

Note. For moderators with more than two levels, mean effect sizes with different subscripts are statistically different from one another, based on a family-wise Type I error probability of .05. CI = confidence interval.

those that were unpublished ($d = -0.33$), $Q(1) = 18.5$, $p < .001$. Design was also statistically significant, with $Q(2) = 26.3$, $p < .001$. Within-subject designs ($d = 1.23$) had higher mean effect sizes than quasi-experimental ($d = 0.27$) and experimental designs ($d = -0.08$). Additionally, implementer was significant, $Q(1) = 23.2$, $p < .001$; teacher-implemented treatments ($d = 0.82$) had greater effect sizes than researcher-delivered programs ($d = -0.39$). The assumption of statistical independence was significant, $Q(1) = 5.5$, $p = .01$. Studies that violated the assumption had greater effect sizes ($d = 0.61$) than those that met the assumption ($d = -0.01$).

Transfer. Table 7 contains the findings for each of the coded moderator variables for the learning outcome of transfer. Level of instructional guidance, $Q(1) = 6.7$, $p = .009$, was statistically significant. In contrast to the findings from the other outcomes, low levels of guidance produced a larger mean effect size ($d =$

0.27) than high levels of guidance ($d = 0.00$). Perceptual richness of the manipulative was also statistically significant, $Q(1) = 12.2$, $p < .001$. Again, differing from the results from the other outcomes, perceptually rich manipulatives had a higher mean effect size ($d = 0.48$) than bland manipulatives ($d = -0.02$) on transfer outcomes.

Two methodological variables were found to significantly moderate effect sizes: the design of the research, $Q(2) = 27.5$, $p < .001$, and the statistical assumption of independence, $Q(1) = 29.3$, $p < .001$. Unlike previous results, significant differences existed between studies in favor of experimental design, with experiments ($d = 0.40$) having a statistically significant higher mean effect size than both within-subjects ($d = 0.25$) and quasi-experiments ($d = -0.21$). Studies that met the assumption of independence had a smaller (and negative) mean effect size ($d = -0.67$) than those that did not meet the assumption ($d = 0.27$).

Table 5
Variability of Effect Sizes Within Retention

Moderator	<i>k</i>	<i>N</i>	<i>d</i>	95% CI	<i>Q</i> _{between}
Methodology characteristics					
Peer review status					90.7, <i>p</i> < .001
Published	23	4,162	0.97	[0.86, 1.07]	
Not published	30	2,978	0.30	[0.21, 0.39]	
Design					183.1, <i>p</i> < .001
Within subject	12	707	1.69 _a	[1.52, 1.87]	
Quasi-experimental	29	5,187	0.35 _b	[0.27, 0.44]	
Experimental	12	1,246	0.47 _b	[0.34, 0.60]	
Implementer					89.1, <i>p</i> < .001
Researcher	15	6,135	0.13	[0.01, 0.24]	
Teacher	38	1,005	0.82	[0.73, 0.90]	
Test type					6.9, <i>p</i> = .008
Standardized	14	1,206	0.77	[0.62, 0.92]	
Researcher created	39	5,934	0.54	[0.47, 0.62]	
Independence					11.9, <i>p</i> < .001
Met	13	937	0.79	[0.66, 0.93]	
Not met	40	6,203	0.52	[0.44, 0.60]	
Instructional characteristics					
Instructional guidance					106.5, <i>p</i> < .001
High	31	4,531	0.90	[0.81, 0.99]	
Low	22	2,609	0.19	[0.08, 0.29]	
Mathematical topic					44.5, <i>p</i> < .001
Place value	3	101	0.70 _{ab}	[0.37, 1.04]	
Arithmetic	25	2,387	0.39 _b	[0.29, 0.48]	
Geometry	5	602	0.57 _{ab}	[0.37, 0.78]	
Fractions	13	2,762	0.93 _a	[0.78, 1.08]	
Algebra	9	1,288	0.84 _a	[0.65, 1.03]	
Perceptual richness					36.4, <i>p</i> < .001
Yes	12	1,395	0.28	[0.14, 0.41]	
No	34	4,723	0.77	[0.69, 0.85]	
Group vs. individual					0.02, <i>p</i> = .87
Individual	5	481	0.57	[0.31, 0.82]	
Group	48	6,659	0.59	[0.52, 0.66]	
Development status					106.8, <i>p</i> < .001
Preoperational	8	707	-0.09 _a	[-0.26, 0.07]	
Concrete	40	6,109	0.81 _b	[0.73, 0.89]	
Formal	5	324	0.31 _c	[0.10, 0.52]	
Instructional time					7.4, <i>p</i> = .02
≤14 days	25	3,133	0.59 _a	[0.49, 0.69]	
15-45 days	15	1,261	0.35 _b	[0.21, 0.49]	
≥46 days	7	952	0.49 _{ab}	[0.28, 0.71]	

Note. For moderators with more than two levels, mean effect sizes with different subscripts are statistically different from one another, based on a family-wise Type I error probability of .05. CI = confidence interval.

Publication Bias

Publication bias refers to the possibility that results from studies showing statistically significant effects in the expected direction are more likely to be published than results from studies not showing such effects. This is commonly referred to as the file-drawer phenomenon (Rosenthal, 1979). To assess the possible impact of such bias, we included both published and unpublished manuscripts in the present review. However, given that our unpublished studies consist primarily of dissertation and thesis projects, we cannot rule out the possibility that other studies with nonsignificant findings have been excluded. Therefore, we conducted an analysis of publication bias to assess the potential impact of missing studies on meta-analytic results. Rosenthal's (1979) fail-safe *N* was calculated to determine how many studies with a null effect would be needed to attenuate the overall effect size to

nonsignificance. This analysis revealed that approximately 9,501 studies would be needed to decrease the average effect size of manipulatives to nonsignificance.

Discussion

The purpose of this meta-analytic review was to examine the effect of using concrete manipulatives for teaching mathematics when compared with abstract symbolic instructional conditions. Additional research exploring comparisons of manipulatives against iconic representations of manipulatives would greatly enhance our understanding of instructional strategies that use concrete representations of mathematics. Currently, the findings from the present review of manipulatives suggest a small- to moderate-sized effect in favor of instructional strategies that use manipulatives when compared with abstract symbolic instruction. However,

Table 6
Variability of Effect Sizes Within Problem Solving

Moderator	<i>k</i>	<i>N</i>	<i>d</i>	95% CI	<i>Q</i> _{between}
Methodology characteristics					
Peer review status					18.5, <i>p</i> < .001
Published	7	335	0.76	[0.50, 1.03]	
Not published	2	142	−0.33	[−0.07, 0.09]	
Design					26.3, <i>p</i> < .001
Within subject	3	146	1.23 _a	[−0.85, 1.60]	
Quasi-experimental	3	144	0.27 _b	[−0.20, 0.75]	
Experimental	3	187	−0.08 _b	[−0.43, 0.25]	
Implementer					23.2, <i>p</i> < .001
Researcher	2	116	−0.39	[−1.85, 0.02]	
Teacher	7	361	0.82	[0.55, 1.09]	
Test type					8.2, <i>p</i> = .002
Standardized	3	201	−0.03	[−0.43, 0.36]	
Researcher created	6	276	0.69	[0.42, 0.97]	
Independence					5.5, <i>p</i> = .01
Met	2	106	−0.01	[−0.47, 0.43]	
Not met	7	371	0.61	[0.35, 0.87]	
Instructional characteristics					
Instructional guidance					19.1, <i>p</i> < .001
High	5	202	1.06	[0.71, 1.42]	
Low	4	275	0.04	[−1.30, 1.39]	
Mathematical topic					45.1, <i>p</i> < .001
Place value	1	24	0.48 _a	[−0.32, 1.29]	
Arithmetic	5	246	0.02 _a	[−0.26, 0.30]	
Geometry	1	93	0.72 _a	[0.20, 1.24]	
Fractions	2	114	2.50 _b	[1.82, 3.18]	
Algebra	0				
Perceptual richness					15.3, <i>p</i> < .001
Yes	2	100	−0.27	[−0.72, 0.18]	
No	6	377	0.80	[0.52, 1.08]	
Group vs. individual					
Individual	0				
Group	9	477	0.46	[0.23, 0.68]	
Development status					2.2, <i>p</i> = .33
Preoperational	1	66	0.08	[−0.58, 0.75]	
Concrete	7	318	0.45	[0.18, 0.72]	
Formal	1	93	0.72	[0.20, 1.24]	
Instructional time					22.7, <i>p</i> < .001
≤14 days	6	290	0.86 _a	[0.56, 1.16]	
15–45 days	1	76	−0.62 _b	[−1.17, −0.62]	
≥46 days	2	111	0.25 _a	[−0.19, 0.69]	

Note. For moderators with more than two levels, mean effect sizes with different subscripts are statistically different from one another, based on a family-wise Type I error probability of .05. CI = confidence interval.

these results cannot be used as evidence that manipulatives are beneficial for learning when making comparisons to other mathematical instructional strategies. Furthermore, an examination of effect sizes across instructional characteristics and learning outcomes revealed that the effectiveness of manipulatives is complex and requires consideration of instructional characteristics and learning outcomes.

Concrete manipulatives have been proposed as an effective strategy in aiding students in problem solving and transfer of mathematical understanding (Burns, 1996; NCTM, 2000). Effect sizes when separated by learning outcome did not follow this assumed pattern. In fact, instruction that used manipulatives produced a moderate- to large-sized effect when students were measured on retention and small effects when higher level outcomes such as problem solving, transfer, and justification were considered. Taken as a whole, the aggregated and disag-

gregated findings indicate that concrete manipulatives may have a differential impact on learning outcomes. These differential outcomes should be considered for future empirical examination and when teaching mathematics concepts with manipulatives.

Instructional Characteristics

Level of instructional guidance, mathematical topic, development status, perceptual richness, and instructional time were statistically significant moderators of the effects of using concrete manipulatives. The moderators of developmental status, level of instructional guidance, type of manipulative, and instructional time are of particular interest due to the connections between these moderators and current developmental and instructional theories.

Table 7
Variability of Effect Sizes Within Transfer

Moderator	<i>k</i>	<i>N</i>	<i>d</i>	95% CI	<i>Q</i> _{between}
Methodology characteristics					
Peer review status					2.2, <i>p</i> = .13
Published	8	2,871	0.19	[0.05, 0.32]	
Not published	5	582	0.04	[−0.15, 0.25]	
Design					27.5, <i>p</i> < .001
Within subject	2	72	0.25 _{ab}	[0.69, 0.12]	
Quasi-experimental	6	702	−0.21 _a	[−0.39, −0.03]	
Experimental	5	2,415	0.40 _b	[0.25, 0.56]	
Implementer					1.4, <i>p</i> = .23
Researcher	4	259	0.06	[−0.20, 0.33]	
Teacher	9	3,194	0.16	[0.04, 0.29]	
Test type					0.92, <i>p</i> = .33
Standardized	1	60	0.18	[−0.43, 0.80]	
Researcher created	12	3,393	0.14	[0.03, 0.26]	
Independence					29.3, <i>p</i> < .001
Met	3	225	−0.61	[−0.91, −0.31]	
Not met	10	3,228	0.27	[0.15, 0.39]	
Instructional characteristics					
Instructional guidance					6.7, <i>p</i> = .009
High	8	2,385	0.00	[−0.16, 0.16]	
Low	5	1,068	0.27	[0.12, 0.43]	
Mathematical topic					1.5, <i>p</i> = .83
Place value	1	12	0.33	[−0.28, 0.96]	
Arithmetic	7	757	0.16	[0.00, 0.36]	
Geometry	1	147	0.24	[−0.15, 0.65]	
Fractions	3	1,996	0.09	[−0.14, 0.32]	
Algebra	1	541	0.09	[−0.18, 0.36]	
Perceptual richness					12.2, <i>p</i> < .001
Yes	5	1,359	0.48	[0.25, 0.62]	
No	6	2,094	−0.02	[−0.23, 0.17]	
Group vs. Individual					
Individual					
Group					
Developmental Status					
Preoperational					
Concrete	13	3,453	0.14	[0.03, 0.26]	
Formal					
Instructional time					3.6, <i>p</i> = .06
≤14 days	7	1,625	0.03	[−0.13, 0.21]	
15–45 days	6	1,828	0.22	[0.07, 0.37]	
≥46 days					

Note. For moderators with more than two levels, mean effect sizes with different subscripts are statistically different from one another, based on a family-wise Type I error probability of .05. CI = confidence interval.

Developmental status. Several contemporary theorists propose that ability to reason abstractly is the pinnacle of cognitive development. According to these theorists, concrete manipulatives should be provided when younger learners begin studying abstract mathematical concepts (Bruner, 1964; Piaget, 1962). Within this developmental framework, it is expected that providing manipulatives allows educators to represent abstract concepts with concrete representations. These concrete representations are expected to facilitate the construction of meaning for pre- and concrete operational children and result in positive cognitive consequences. In contrast, students who are assumed to have reached formal operations are not expected to derive comparable cognitive benefits from the provision of concrete manipulatives.

Our findings provide partial support for these developmental predictions. At the aggregated level, studies that included children

assumed to have facility with concrete operations showed medium to large effect sizes, whereas studies comprising formal operational students had relatively smaller effect sizes. Within the learning outcome of retention, the pattern of manipulatives being the most efficacious for students within the assumed concrete operations stage remains; however, studies consisting of samples of preoperational-age children revealed a statistically lower and negative mean effect size ($d = -0.09$) than studies consisting of samples of assumedly concrete or formal operational students. Recently, developmental theorists have proposed explanations for why concrete manipulatives may be less effective with younger children. According to these explanations, concrete manipulatives may not be as effective with younger children because they may struggle with the concept that an object can stand for the item while simultaneously representing a larger mathematical concept (DeLoache, 2000; Uttal et al., 2009). Future research should ex-

amine how concrete manipulatives assist in developing foundational mathematical concepts with younger children.

Instructional guidance. Conflicting recommendations have been provided to practitioners concerning the level of instructional guidance offered to students during the learning process. Those who recommend high levels of instructional guidance propose that instructional guidance provides students with explicit opportunities to select pertinent information, organize the information into coherent structures, and incorporate the new information with prior knowledge (Mayer, 2003). According to this model, low levels of instructional guidance do not promote this process due to the lack of explicit guidance selecting relevant information. Results from the aggregated, retention, and problem-solving data support this model, with high levels of guidance being associated with higher levels of student learning. This finding also aligns with prior research examining the efficacy of manipulatives in listening and reading instruction (Glenberg, Brown, & Levin, 2007; Glenberg, Jaworski, Rischal, & Levin, 2007; Marley et al., 2007). These researchers have suggested that the scripted manipulation of objects helps students establish connections between concrete representations and their abstract referents (i.e., words), which in turn enhances comprehension.

Proponents of low instructional guidance contend that students who reach proficiency with limited or no instructional guidance develop greater conceptual understandings and are subsequently more adept at transferring this knowledge to novel circumstances (Schauble, 1996; Stohr-Hunt, 1996). Martin's (2009) theory of physically distributed learning supports this notion with the explanation that students are able to impose their own meanings on manipulatives. This development of self-relevant meaning allows for greater flexibility and for learning to be transferred to novel circumstances. The moderator analysis associated with the transfer of learning outcome provides partial support for this perspective, with low instructional guidance studies having larger effect sizes on transfer of learning relative to high levels of guidance. Future research is needed to better understand what level of instructional guidance is optimum for student learning with manipulatives. More specifically, research is needed to examine how level of instructional guidance may need to vary depending upon learning objective.

Perceptual richness. The perceptual richness of manipulatives has been identified as a potential deterrent to student learning and performance (Kaminski et al., 2009; Martin & Schwartz, 2005; McNeil et al., 2009). Superficial details that are present in perceptually rich manipulatives have been shown to distract children when asked to perform a math word problem, resulting in students making more errors in solving the math problem but proportionally fewer conceptual errors than students who used bland manipulatives (McNeil et al., 2009).

The moderator analysis within the retention and problem-solving data provides additional support for the idea that perceptually rich manipulatives suppress student learning. Within the learning outcome of retention, studies utilizing perceptually rich manipulatives had a smaller effect on student measures of immediate performance. Additionally, specific to the learning outcome of problem solving, those studies that used perceptually rich manipulatives revealed a lower and negative mean effect size ($d = -0.27$) when compared with studies that used bland manipulatives during the problem solving process.

Results on transfer of learning, an outcome that requires greater conceptual understanding of the mathematics concepts, indicated that perceptually rich manipulatives may enhance student learning. With findings indicating that studies that used bland manipulatives had a lower and negative mean effect size ($d = -0.02$) than studies using perceptually rich manipulatives. However, it is important to note that this finding contradicts previous cognitive research that suggests that the perceptual richness of images inhibits the transfer of learning (Goldstone & Sakamoto, 2003; Kaminski, Sloutsky, & Heckler, 2008). To better understand this finding, additional research examining the relationship between the perceptual richness of manipulatives and their effects on a variety of student learning outcomes is warranted.

Instructional time. The length of instructional time provided to students has been established as an essential variable to learning (Rosenshine & Berliner, 1978). Furthermore, experiments contrasting direct and discovery learning as means to improving inquiry skills have implicated length of instructional time as an explanation for inconsistent empirical results (Dean & Kuhn, 2007) within the instructional strategy literature. According to Dean and Kuhn (2007), in order for student-controlled strategies to be effective, students must engage in instruction over an extended period. In addition, Sowell's (1989) meta-analysis provides evidence that extended use of manipulatives had a positive effect on measures of retention. Results from the moderator analysis of the present study contradict these findings. Studies that were less than 45 days had a higher mean effect on student learning within the aggregated data. Additionally, within the learning outcomes of retention, studies that were less than 14 days had a higher mean effect than studies that were longer (i.e., more than 15 days), and within problem solving, studies that were coded at being medium in length (15 to 45 days) had a lower and negative mean effect size ($d = -0.62$) than short or long studies. A possible explanation for these contradictory findings was the inability of the coding to disentangle instructional time with the length of study. Therefore, further research that specifically examines varying lengths of instructional time with manipulatives is needed to provide a better understanding of how this instructional variable moderates the overall effectiveness of manipulatives.

Methodological Characteristics

Methodological aspects of studies presented additional variation in the overall findings. This variation in effect sizes due to methodological characteristics is related to the validity of the inferences made from the results of this literature. Methodological characteristics that significantly moderated effect sizes included peer review status, research design, implementer, type of test, and accounting for the statistical assumption of independence. Moderation within these variables raises specific concerns related to statistical conclusion validity and internal validity.

Statistical conclusion validity. Validity concerning the inferences made about the covariation between treatment and outcome is of specific concern when examining the efficacy of teaching strategies. Threats to statistical conclusion validity may lead to results that either overestimate or underestimate the magnitude of the covariation between the treatment and the outcome (Shadish et al., 2002). When threats of this nature have a high prevalence in the literature, conclusions regarding the effectiveness of the teach-

ing strategy are limited. This concern is supported by the findings that a greater effect size was produced when statistical independence is not accounted for in the analysis. Furthermore, published studies and within-subjects studies produced larger effect sizes. The latter suggests that the results of within-subjects studies should be carefully evaluated. The former could be an indication that published studies are of higher quality and produce significant results, or that significant findings tend to be published over findings that fail to reach statistical significance.

Internal validity. The ability to make an inference about the causal effect of manipulatives on student learning is grounded in research design. At a minimum, to build a case for strong internal validity, studies must be able to account for plausible rival explanations. The most effective way to control for threats to internal validity is the use of an experimental design. By randomly assigning participants to conditions, plausible threats to internal validity are minimized. The finding that studies that did not use random assignment had a significantly higher effect size than those from experimental designs emphasizes the need for researchers and practitioners to be cautious when making prescriptive statements concerning the efficacy of manipulatives.

Limitations

Limitations of meta-analyses are often similar to those for primary studies (Card, 2012). Three such limitations common to both meta-analytical and primary studies that are relevant to the current meta-analysis are the potential problem of generalizing results, issues related to incomplete or missing data, and the difficulty of drawing sound inference about causal effects.

Specific to this meta-analysis is the possible limitation of generalizability due to our definitions of manipulative and control groups. For example, studies that used a scale were not included in the sample, nor were studies that made comparisons between manipulatives and pictures. Inclusion of studies that used tools or pictures may alter conclusions regarding mathematics manipulatives. Therefore, the circumscribed conclusion that can be drawn is that a positive effect of manipulatives was found on student learning outcomes when it was compared with conditions in which no manipulatives or other concrete materials (e.g., pictures) were used and that the relationship between manipulatives and learning within this comparison is moderated by different instructional characteristics. Further research should be conducted to collect evidence comparing manipulatives to other learning strategies.

Another potential limitation on generalizability is posed by the previously mentioned file-drawer problem. A careful attempt was made to exhaustively search for unpublished manuscripts; however, some studies examining the efficacy of manipulatives may not have been identified. This means the possibility of publication bias is still present. Additionally, in many instances, studies included in the meta-analysis did not report enough information for all variables of interest to be coded. The lack of information from individual studies created numerous occasions of missing values for the moderator variables. The inability to use all studies in each moderator analysis could have decreased the statistical power of the meta-analysis to detect differences that may exist. The lack of information from studies also created many missed opportunities to further develop an understanding of the efficacy of manipulatives. Detailed information related to the procedure of the study

(i.e., interval of time between outcomes measures, the use of manipulatives at time of testing, and details related to actual activities of participants) may have yielded a slightly different overall outcome on student learning. For example, the lack of details provided in articles limited the ability to make distinctions between learning activities that were conducted within each study. Clarification of the activities conducted within each study may have provided an opportunity to use established frameworks such as Chi's (2009) taxonomy of learning activities.

Finally, as noted by Card (2012), the strength of conclusions from any meta-analysis is based on the quality of the research design for both the meta-analysis and each of the constituent studies. For example, results from a meta-analysis of high-quality studies will always yield better conclusions than a meta-analysis of low-quality studies. Therefore, the observed methodological differences in study quality may affect the quality of the present results. In addition, it should be emphasized that the reported moderator effects are based on the observed covariation between the coded moderator variables and the values of the study effect sizes. As such, moderator effects should not be construed as strong evidence for the causal effect of a moderator variable on the effectiveness of using manipulative.

Conclusion

Results from this meta-analysis begin to focus the inconsistencies seen within the manipulation-based literature. Findings indicate that using manipulatives in mathematics instruction produces a small- to medium-sized effect on student learning when compared with instruction that uses abstract symbols alone. Additionally, results revealed that the strength of this effect is dependent upon other instructional variables. Instructional variables such as the perceptual richness of an object, level of guidance offered to students during the learning process, and the development status of the learner moderate the efficacy of manipulatives. The finding that specific instructional variables either suppress or increase the efficacy of manipulatives suggests that simply incorporating manipulatives into mathematics instruction may not be enough to increase student achievement in mathematics. These contextual variables therefore must be considered when planning instruction. It is our hope that the results of this meta-analysis will further stimulate math manipulatives research and assist others in generating new and more specific hypotheses investigating this instructional strategy.

References

References marked with an asterisk indicate studies included in the meta-analysis.

- *Aburime, E. F. (2007). How manipulatives affect the mathematics achievement of students in Nigerian schools. *Education Research Quarterly*, 3(1), 3–16.
- Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1–18. doi:10.1037/a0021017
- *Anderson, G. R. (1957). Visual-tactual devices and their efficacy: An experiment in grade eight. *Arithmetic Teacher*, 4(3), 196–203.
- *Aurich, M. R. (1963). *A comparative study to determine the effectiveness of the Cuisenaire method of arithmetic instruction with children at the*

- first grade level (Unpublished doctoral dissertation). Catholic University of America, Washington, DC.
- *Babb, J. H. (1975). The effects of textbook instruction, manipulatives and imagery on recall of the basic multiplication facts. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 36, 4378.
- Baranes, R., Perry, M., & Stigler, J. W. (1989). Activation of real-world knowledge in the solution of word problems. *Cognition and Instruction*, 6(4), 287–318. doi:10.1207/s1532690xci0604_1
- *Battle, T. S. (2007). *Infusing math manipulatives: The key to an increase in academic achievement in the mathematics classroom*. (ERIC Document Reproduction Service No. ED498579)
- Biazak, J. E., Marley, S. C., & Levin, J. R. (2010). Does an activity-based learning strategy improve preschool children's memory for narrative passages? *Early Childhood Research Quarterly*, 25(4), 515–526. doi:10.1016/j.ecresq.2010.03.006
- Billstein, R., Libeskind, S., & Lott, J. W. (2009). *A problem solving approach to mathematics: For elementary school teachers*. Boston, MA: Addison Wesley.
- *Bring, C. R. (1972). Effects of varying concrete activities on the achievement of objectives in metric and non-metric geometry by students of grades five and six. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 32(7), 3775.
- Brown, M. C., McNeil, N. M., & Glenberg, A. M. (2009). Using concreteness in education: Real problems, potential solutions. *Child Development Perspectives*, 3(3), 160–164. doi:10.1111/j.1750-8606.2009.00098.x
- Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, 31, 21–32.
- Bruner, J. S. (1964). The course of cognitive growth. *American Psychologist*, 19(1), 1–15. doi:10.1037/h0044160
- Burns, M. (1996). How to make the most of math manipulatives. *Instructor*, 105(7), 45–51.
- *Butler, F. M., Miller, S. P., Crehan, K., Babbitt, B., & Pierce, T. (2003). Fraction instruction for students with mathematics disabilities: Comparing two teaching sequences. *Learning Disabilities Research & Practice*, 18(2), 99–111. doi:10.1111/1540-5826.00066
- Canobi, K. H. (2005). Children's profiles of addition and subtraction understanding. *Journal of Experimental Child Psychology*, 92(3), 220–246. doi:10.1016/j.jecp.2005.06.001
- Carbonneau, K. J., & Marley, S. C. (in press). Activity-based learning strategies and academic achievement. In J. A. C. Hattie & E. M. Anderman (Eds.), *The international handbook of student achievement*. New York, NY: Routledge.
- Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford Press.
- *Carmody, L. M. (1970). A theoretical and experimental investigation into the role of concrete and semi-concrete materials in the teaching of elementary school mathematics. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 31, 3407.
- Chi, M. T. H. (2009). Active–constructive–interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1), 73–105. doi:10.1111/j.1756-8765.2008.01005.x
- Clark, J. M., & Paivio, A. (1987). A dual coding perspective on encoding processes. In M. A. McDaniel & M. Pressley (Eds.), *Imagery and related mnemonic processes: Theories, individual differences, and applications* (pp. 5–33). New York, NY: Springer-Verlag.
- Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- *Cook, D. M. (1967). *Research and development activities in R & I units of two elementary schools in Janesville, Wisconsin, 1966–1967 (Report No. TR-45)*. Washington, DC: U.S. Department of Health, Education and Welfare.
- Copley, J. V. (2000). *The young child and mathematics*. Washington, DC: National Association for the Education of Young Children. (ERIC Document Reproduction Service No. ED448906)
- *Cramer, K. A., Post, T. R., & delMas, R. C. (2002). Initial fraction learning by fourth- and fifth-grade students: A comparison of the effects of using commercial curricula with the effects of using the Rational Number Project Curriculum. *Journal for Research in Mathematics Education*, 33(2), 111–144. doi:10.2307/749646
- *Dawson, D. T. (1955). An experimental approach to the division idea. *Arithmetic Teacher*, 2(1), 6–9.
- Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, 91(3), 384–397. doi:10.1002/sce.20194
- DeLoache, J. S. (2000). Dual representation and young children's use of scale models. *Child Development*, 71(2), 329–338. doi:10.1111/1467-8624.00148
- *Dyer, L. (1996). *An investigation of the use of algebraic manipulatives with community college students* (Unpublished doctoral dissertation). University of Missouri, Columbia, MO.
- *Egan, D. (1990). *The effects of using Cuisenaire rods on math achievement of second grade students* (Unpublished doctoral dissertation). Central Missouri State University, Warrensburg, MO.
- *Ekman, L. G. (1967). A comparison of the effectiveness of different approaches to the teaching of addition and subtraction algorithms in the third grade. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 27, 2275–2276.
- Engelkamp, J., Zimmer, H. D., Mohr, G., & Sellen, O. (1994). Memory of self-performed tasks: Self-performing during recognition. *Memory & Cognition*, 22(1), 34–39. doi:10.3758/BF03202759
- Fennema, E. H. (1972). The relative effectiveness of a symbolic and a concrete model in learning a selected mathematical principle. *Journal for Research in Mathematics Education*, 3(4), 233–238. doi:10.2307/748490
- *Fujimura, N. (2001). Facilitating children's proportional reasoning: A model of reasoning processes and effects of intervention on strategy change. *Journal of Educational Psychology*, 93(3), 589–603. doi:10.1037/0022-0663.93.3.589
- *Garcia, E. P. (2004). *Using manipulatives and visual cues with explicit vocabulary enhancement for mathematics instruction with grade three and four low achievers in bilingual classrooms* (Unpublished doctoral dissertation). Texas A&M University, College Station, TX.
- *Getgood, J. F. (2000). *The effect of factor blocks, a manipulative, in student understanding of greatest common factor, least common multiple and prime factorization* (Unpublished doctoral dissertation). George Mason University, Fairfax, VA.
- Glenberg, A. M., Brown, M., & Levin, J. R. (2007). Enhancing comprehension in small reading groups using a manipulation strategy. *Contemporary Educational Psychology*, 32(3), 389–399. doi:10.1016/j.cedpsych.2006.03.001
- Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and imagined activity can enhance young children's reading comprehension. *Journal of Educational Psychology*, 96(3), 424–436. doi:10.1037/0022-0663.96.3.424
- Glenberg, A. M., Jaworski, B., Rischal, M., & Levin, J. (2007). What brains are for: Action, meaning, and reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 221–240). Mahwah, NJ: Erlbaum
- *Goins, K. B. (2001). Comparing the effects of visual and algebra tile manipulative methods on student skill and understanding of polynomial multiplication. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 62, 12.
- Goldstone, R. L., & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology*, 46(4), 414–466. doi:10.1016/S0010-0285(02)00519-4

- *Gürbüüz, R. (2010). The effect of activity-based instruction on conceptual development of seventh grade students in probability. *International Journal of Mathematical Education in Science and Technology*, 41(6), 743–767. doi:10.1080/00207391003675158
- *Hawkins, V. J. (1982). A comparison of two methods of instruction, a saturated learning environment and traditional learning environment: Its effects on achievement and retention among female adolescents in first-year algebra. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 43, 02.
- Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, 7(2), 119–137. doi:10.2307/1164961
- Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, 93(2), 388–395. doi:10.1037/0033-2909.93.2.388
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- *Hiebert, J., Wearne, D., & Taber, S. (1991). Fourth graders' gradual construction of decimal fractions during instruction using different physical representations. *Elementary School Journal*, 91(4), 321–341. doi:10.1086/461658
- *Hinzman, K. P. (1997). *Use of manipulatives in mathematics at the middle school level and their effects on student grades and attitudes* (ERIC Document Reproduction Service No. ED411150).
- *Johnson, R. L. (1970). *Effects of varying concrete activities on achievement of objectives in perimeter, area and volume by students grades four five and six* (Unpublished doctoral dissertation). University of Colorado, Denver, CO.
- *Jordan, L., Miller, M. D., & Mercer, C. D. (1999). The effects of concrete to semiconcrete to abstract instruction in the acquisition and retention of fraction concepts and skills. *Learning Disabilities*, 9(3), 115–122.
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. (2008). The advantage of abstract examples in learning math. *Science*, 320(5875), 454–455. doi:10.1126/science.1154659
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. (2009). Transfer of mathematical knowledge: The portability of generic instantiations. *Child Development Perspectives*, 3(3), 151–155. doi:10.1111/j.1750-8606.2009.00096.x
- King, V. C. (1976). A study of the effect of selected mathematics instructional sequences on retention and transfer with logical reasoning controlled. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 37(5), 2696–2697.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86. doi:10.1207/s15326985ep4102_1
- Kormi-Nouri, R., Nyberg, L., & Nilsson, L. G. (1994). The effect of retrieval enactment on recall of subject-performed tasks and verbal tasks. *Memory & Cognition*, 22(6), 723–728. doi:10.3758/BF03209257
- *Kuhfittig, P. K. F. (1974). The relative effectiveness of concrete aids in discovery learning. *School Science and Mathematics*, 74(2), 104–108. doi:10.1111/j.1949-8594.1974.tb09207.x
- *LeBlanc, J. F. (1968). The performance of first grade children in four levels of conservation of numerosness and three IQ groups when solving arithmetic subtraction problems. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 29, 67A.
- Lefrançois, G. R. (1997). *Psychology for teachers*. Belmont, CA: Wadsworth.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- *Lucas, J. S. (1966). The effects of attribute block training on children's development of arithmetic concepts. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 27, 2400–2401.
- *Lucow, W. H. (1964). An experiment with the Cuisenaire method in grade three. *American Educational Research Journal*, 1(3), 159–167. doi:10.2307/1162217
- Marley, S. C., & Levin, J. R. (2006). Pictorial illustrations, visual imagery, and motor activity: Their instructional implications for Native American children with learning disabilities. In R. J. Morris (Ed.), *Disability research and policy: Current perspectives* (pp. 103–123). Mahwah, NJ: Erlbaum.
- Marley, S. C., & Levin, J. R. (2011). When are prescriptive statements in educational research justified? *Educational Psychology Review*, 23(2), 197–206. doi:10.1007/s10648-011-9154-y
- Marley, S. C., Levin, J. R., & Glenberg, A. M. (2007). Improving Native American children's listening comprehension through concrete representations. *Contemporary educational psychology*, 32(3), 537–550. doi:10.1016/j.cedpsych.2007.03.003
- Marley, S. C., Levin, J. R., & Glenberg, A. M. (2010). What cognitive benefits does an activity-based reading strategy afford young Native American readers? *Journal of Experimental Education*, 78(3), 395–417. doi:10.1080/00220970903548061
- Marley, S. C., & Szabo, Z. (2010). Improving children's listening comprehension with a manipulation strategy. *Journal of Educational Research*, 103(4), 227–238. doi:10.1080/00220670903383036
- Marley, S. C., Szabo, Z., Levin, J. R., & Glenberg, A. M. (2011). Investigation of an activity-based text-processing strategy in mixed-age child dyads. *Journal of Experimental Education*, 79(3), 340–360. doi:10.1080/00220973.2010.483697
- Martin, T. (2009). A theory of physically distributed learning: How external environments and internal states interact in mathematics learning. *Child Development Perspectives*, 3(3), 140–144. doi:10.1111/j.1750-8606.2009.00094.x
- Martin, T., & Schwartz, D. L. (2005). Physically distributed learning: Adapting and reinterpreting physical environments in the development of fraction concepts. *Cognitive Science*, 29(4), 587–625. doi:10.1207/s15516709cog0000_15
- Mayer, R. E. (2003). *Learning and instruction*. Upper Saddle River, NJ: Prentice Hall.
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, 59(1), 14–19. doi:10.1037/0003-066X.59.1.14
- *McClung, L. W. (1998). *A study on the use of manipulatives and their effect on student achievement in high school Algebra I class*. (ERIC Document Reproduction Service No. ED425077)
- McNeil, N., & Jarvin, L. (2007). When theories don't add up: disentangling the manipulatives debate. *Theory Into Practice*, 46(4), 309–316. doi:10.1080/00405840701593899
- McNeil, N. M., Uttal, D. H., Jarvin, L., & Sternberg, R. J. (2009). Should you show me the money? Concrete objects both hurt and help performance on mathematics problems. *Learning and Instruction*, 19(2), 171–184. doi:10.1016/j.learninstruc.2008.03.005
- *Miller, J. W. (1964). An experimental comparison of two approaches to teaching multiplication of fractions. *Journal of Educational Research*, 57(9), 468–471.
- Montessori, M. (1964). *The Montessori method*. New York, NY: Schocken.
- *Moody, W. B., Abell, R., & Bausell, R. B. (1971). The effect of activity-oriented instruction upon original learning, transfer, and retention. *Journal for Research in Mathematics Education*, 2(3), 207–212. doi:10.2307/749045
- *Nasca, D. (1966). Comparative merits of a manipulative approach to second-grade arithmetic. *Arithmetic Teacher*, 13(3), 221–225.
- National Center for Education Statistics. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context*. Washington, DC: U.S. De-

- partment of Education. Retrieved from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009001>
- National Center for Education Statistics. (2011). *The nation's report card: Mathematics 2011*. Washington, DC: U.S. Department of Education. Retrieved from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012458>
- National Council of Teachers of Mathematics. (2000). *Standards & focal points*. Retrieved from <http://www.nctm.org/standards/default.aspx?id=58>
- *Nichols, E. J. (1972). A comparison of two methods of instruction in multiplication and division for third grade pupils. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 32, 6011.
- *Nickel A. P. (1971). A multi-experience approach to conceptualization for the purpose of improvement of verbal problem-solving in arithmetic. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 32, 2917–2918.
- *Nishida, T. K. (2007). The use of manipulatives to support children's acquisition of abstract math concepts. *Dissertation Abstracts International: Section B. Sciences and Engineering*, 69(1), 718.
- *Norman, M. (1955). *Three methods of teaching basic division facts* (Unpublished doctoral dissertation). State University of Iowa, Ames, IA.
- *Olkun, S. (2003). Comparing computer versus concrete manipulatives in learning 2D geometry. *Journal of Computers in Mathematics and Science Teaching*, 22(1), 43–56.
- Paivio, A. (1986). *Mental representations: A dual coding approach*. New York, NY: Oxford University Press.
- *Paolini, M. W. (1977). The use of manipulative materials versus non-manipulative materials in a kindergarten mathematics program. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 39, 5382.
- Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York, NY: Basic Books.
- *Peterson, S. K., Mercer, C. D., & O'Shea, L. (1988). Teaching learning disabled students place value using the concrete to abstract sequence. *Learning Disabilities Research*, 4(1), 52–56.
- Piaget, J. (1962). *Play, dreams, and imitation in childhood*. New York, NY: Norton.
- Piaget, J., & Colman, D. (1974). *Science of education and the psychology of the child*. New York, NY: Grossman.
- *Prigge, G. R. (1978). The differential effects of the use of manipulative aids on the learning of geometric concepts by elementary school children. *Journal for Research in Mathematics Education*, 9(5), 361–367. doi:10.2307/748772
- Resnick, L. B., & Omanson, S. F. (1987). Learning to understand arithmetic. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 3, pp. 41–95). Hillsdale, NJ: Erlbaum.
- Rittle-Johnson, B., & Koedinger, K. R. (2005). Designing knowledge scaffolds to support mathematical problem solving. *Cognition and Instruction*, 23(3), 313–349. doi:10.1207/s1532690xc2303_1
- *Robinson, E. B. (1978). *The effects of a concrete manipulative on attitude toward mathematics and levels of achievement and retention of a mathematical concept among elementary students* (Unpublished doctoral dissertation). East Texas State University, Commerce, TX.
- Rosenshine, B. V., & Berliner, D. C. (1978). Academic engaged time. *British Journal of Teacher Education*, 4(1), 3–16. doi:10.1080/0260747780040102
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641. doi:10.1037/0033-2909.86.3.638
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
- Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, 99(3), 400–406. doi:10.1037/0033-2909.99.3.400
- Sarama, J., & Clements, D. H. (2009). "Concrete" computer manipulatives in mathematics education. *Child Development Perspectives*, 3(3), 145–150. doi:10.1111/j.1750-8606.2009.00095.x
- Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, 32(1), 102–119. doi:10.1037/0012-1649.32.1.102
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- *Sherman, J., & Bisanz, J. (2009). Equivalence in symbolic and nonsymbolic contexts: Benefits of solving problems with manipulatives. *Journal of Educational Psychology*, 101(1), 88–100. doi:10.1037/a0013156
- *Shocraft, P. J. (1971). *The effects of provisions for imagery through materials and drawings on translating algebra word problems*. (ERIC Document Reproduction Service No. ED071857)
- *Slaughter, H. B. (1980, April). *Using Title I control group for evaluation research of a supplemental mathematics project for third and fifth grade students*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.
- *Smith, L. F., & Montani, T. O. (2008). The effects of instructional consistency: Using manipulatives and teaching strategies to support resource room mathematics instructions. *Learning Disabilities*, 15(2), 71–76.
- Sowell, E. J. (1989). Effects of manipulative materials in mathematics instruction. *Journal for Research in Mathematics Education*, 20(5), 498–505. doi:10.2307/749423
- *Steen, K., Brooks, D., & Lyon, T. (2006). The impact of virtual manipulatives on first grade geometry instruction and learning. *Journal of Computers in Mathematics and Science Teaching*, 25(4), 373–391.
- *Steger, C. (1977). The effects of two classes of sensory stimuli and race on the acquisition and transfer of the mathematics principle place value. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 37, 4120.
- Stohr-Hunt, P. M. (1996). An analysis of frequency of hands-on experience and science achievement. *Journal of Research in Science Teaching*, 33(1), 101–109. doi:10.1002/(SICI)1098-2736(199601)33:1<101::AID-TEA6>3.0.CO;2-Z
- *Suh, J., & Moyer, P. (2007). Developing students' representational fluency using virtual and physical algebra balances. *Journal of Computers in Mathematics and Science Teaching*, 26(2), 155–173.
- *Taylor, F. M. (2001). *Effectiveness of concrete and computer simulated manipulatives on elementary students learning skills and concepts in experimental probability* (Unpublished doctoral dissertation). University of Florida, Gainesville, FL.
- *Threadgill-Sowder, J. A., & Juilfs, P. A. (1980). Manipulative versus symbolic approaches to teaching logical connectives in junior high school: An Aptitude \times Treatment interaction study. *Journal for Research in Mathematics Education*, 11(5), 367–374. doi:10.2307/748627
- Tindall-Ford, S., & Sweller, J. (2006). Altering the modality of instructions to facilitate imagination: Interactions between the modality and imagination effects. *Instructional Science*, 34(4), 343–365. doi:10.1007/s11251-005-6075-5
- Uttal, D. H., O'Doherty, K., Newland, R., Hand, L. L., & DeLoache, J. (2009). Dual representation and the linking of concrete and symbolic representations. *Child Development Perspectives*, 3(3), 156–159. doi:10.1111/j.1750-8606.2009.00097.x
- *Wallace, P. (1974). An investigation of the relative effects of teaching a mathematical concept via multisensory models in elementary school mathematics. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 35, 2998–2999.
- *Weber, A. W. (1970). Introducing mathematics to first grade children: Manipulative vs. paper/pencil. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*, 30, 3373–3374.

- The White House, Office of the Press Secretary. (2009). *President Obama launches "Educate to Innovate" campaign for excellence in science, technology, engineering & math (STEM) education* [Press release]. Retrieved from <http://www.whitehouse.gov/the-press-office/president-obama-launches-educate-innovate-campaign-excellence-science-technology-en>
- *Witzel, B. S., Mercer, C. D., & Miller, D. (2003). Teaching algebra to students with learning difficulties: An investigation of an explicit instruction model. *Learning Disabilities Research & Practice, 18*(2), 121–131. doi:10.1111/1540-5826.00068
- *Wood, C. M. (1974). *A comparison of the effects of sequence and mode upon the initial acquisition, retention, and transfer of elementary multiplication concepts*. (ERIC Document Reproduction Service No. ED108909)
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Allen, TX: DLM.
- *Yuan, Y., Lee, C.-Y., & Wang, C.-H. (2010). A comparison study of polyominoes explorations in a physical and virtual manipulative environment. *Journal of Computer Assisted Learning, 26*(4), 307–316. doi:10.1111/j.1365-2729.2010.00352.x

Received January 4, 2012

Revision received October 16, 2012

Accepted November 1, 2012 ■

Modeling Writing Development: Contribution of Transcription and Self-Regulation to Portuguese Students' Text Generation Quality

Teresa Limpo and Rui A. Alves
University of Porto

Writing is a complex activity that requires transcription and self-regulation. We used multiple-group structural equation modeling to test the contribution of transcription (handwriting and spelling), planning, revision, and self-efficacy to writing quality at 2 developmental points (Grades 4–6 vs. 7–9). In Grades 4–6, the model explained 76% of the variance in writing quality, and transcription contributed directly to text generation. This finding suggests that, for younger students, handwriting and spelling were the strongest constraints to text generation. In Grades 7–9, the model explained 82% of the variance in writing quality. Although transcription did not contribute directly to text generation, it contributed indirectly through planning and self-efficacy. The progressive automatization of transcription throughout school years may contribute to the acquisition and development of self-regulatory skills, which, in turn, positively influence the quality of text generation. Explicit instruction and practice in handwriting, spelling, planning, and revising along with nurturing of realistic self-efficacy beliefs may facilitate writing development beyond primary years of schooling.

Keywords: writing development, transcription, planning, revision, self-efficacy

From a cognitive perspective, writing is a complex and costly skill that places multiple demands on the writer (Hayes, 1996). Writing is such a complex and demanding activity that it generally takes more than two decades to achieve writing expertise (Kellogg, 2008). Berninger and colleagues have argued that both the simple view of writing proposed by Juel, Griffith, and Gough (1986; Juel, 1988) and the not-so-simple view of writing (Berninger & Winn, 2006; see also Berninger & Chanquoy, 2012) contribute to better understand the writing processes and how they may change over development. In the not-so-simple model, text generation is supported by the collaboration between transcription (handwriting and spelling) and high-level cognitive skills for self-regulation, such as planning and revising. During writing, the interaction among these processes occurs within working memory constraints. In a notable review, Graham and Harris (2000) also concluded that writing development depends on the automatization of transcription and the acquisition of high levels of self-regulation.

The present study aims to contribute to extant research on writing development by focusing on the role of transcription and

self-regulation skills in writing. Although considerable research has shown that these skills influence writing quality, little is known about their relative contribution to text generation throughout schooling. Moreover, studies have been yielding contradictory findings regarding the relationships between transcription and self-regulation and their contribution to written composition from a developmental perspective. The current study was therefore designed to examine the relationships among transcription, self-regulation, and text generation and to directly compare them at two developmental points (Grades 4–6 [age 9–12 years] vs. Grades 7–9 [age 12–15], with about 60 children per grade level). To our knowledge, no such large and comprehensive assessment study, using multiple-group structural equation modeling, has investigated the joint development of these critical writing skills across 6 years of schooling.

Transcription Predicts Writing Quality

Transcription refers to the transformation of language representations in working memory into written text (Berninger, 1999; Graham, Berninger, Abbott, Abbott, & Whitaker, 1997). This requires the retrieval of orthographic symbols and the execution of fine-motor movements for producing them (Abbott & Berninger, 1993). Thus, transcription involves spelling and handwriting.

This low-level writing skill was under-recognized for years (Medwell & Wray, 2008) because it was assumed that it did not interfere with text quality in typically developing children beyond primary grades (Scardamalia, Bereiter, & Goleman, 1982). Nevertheless, during the last two decades, writing research has been accumulating evidence about the impact of transcription in the quality of texts produced by children and adolescents, with and without disabilities (Connelly, Gee, & Walsh, 2007; De La Paz & Graham, 1995; Graham, 1990; MacArthur & Graham, 1987; Reece & Cumming, 1996). Graham et al. (1997; see also Graham &

This article was published Online First February 4, 2013.

Teresa Limpo and Rui A. Alves, Faculty of Psychology and Educational Sciences, University of Porto, Porto, Portugal.

The study reported in this article was supported by Portuguese Foundation for Science and Technology Grant SFRH/BD/68548/2010 to Teresa Limpo and Grant PTDC/PSI-PCO/I10708/2009 to Rui A. Alves. We thank Raquel Fidalgo, David Dickinson, and Amanda Goodwin for helpful discussions about the design of the study and Marcos Dias for helpful comments on earlier versions of the article.

Correspondence concerning this article should be addressed to Teresa Limpo, Faculdade de Psicologia e de Ciências da Educação, Universidade do Porto, Rua Alfredo Allen, 4200-392 Porto, Portugal. E-mail: tlimpo@fpce.up.pt

Harris, 2000) reviewed several correlational studies and concluded that transcription was moderately correlated with text quality. However, this finding should be read carefully as, in the majority of these studies, spelling and handwriting bias were not removed from text quality scoring. This is problematic because it was observed that poor spelling and penmanship have a negative impact on holistic assessments of text quality (Berninger & Swanson, 1994). In the studies reviewed next, this methodological limitation was addressed by setting apart transcription skills from quality assessments.

Regarding spelling, Juel (1988) found that, in Grade 1, 29% of the variance in writing quality was explained by spelling skills, but in Grade 4 the explained variance dropped to 10%. In a 5-year longitudinal study (Grades 1–7), Abbott, Berninger, and Fayol (2010) found that spelling was the most consistent predictor of composing across adjacent grades ($.25 < \beta < .67$). Using structural equation modeling with multiple measures of each construct, Graham et al. (1997) showed that handwriting fluency contributed to writing quality in Grades 1–3 ($\beta = .53$) as much as in Grades 4–6 ($\beta = .67$). Alves and Jesus (2011) found significant correlations between handwriting fluency and writing quality in Grade 2 ($r = .36$), but not in Grades 1, 3, and 4. Christensen (2004) found moderate correlations with a sample of older students (Grades 8–9; $r = .44$). Generally, these studies have shown that writing quality is influenced by writers' transcription skills, even though results are mixed concerning the developmental pattern of this relationship. This might be due in part to whether single or multiple measures were used to assess handwriting fluency, spelling, and compositional quality and also to whether cross-sectional or longitudinal research designs were used.

Berninger and colleagues conducted a comprehensive cross-sectional study collecting multiple transcription and text generation measures from Grade 1 to 9 (for reviews see Berninger & Swanson, 1994; Berninger, 1999). They found that in Grades 1–3 (age 6–9) and Grades 4–6 (age 9–12), respectively, 25% and 42% of the variance in compositional quality was explained by transcription (see also Graham et al., 1997). It is noteworthy that the explained variance in writing quality by transcription dropped to 18% in Grades 7–9 (age 12–15). Although this decrease was not statistically tested, it was suggested that students became more proficient in transcription, and these processes may have exerted less constraint on text generation (Berninger, 1999).

Self-Regulation Predicts Writing Quality

Self-regulation is critical in writing as it enables writers to attain their literary goals through the use of strategies employed before, during, and after writing (Zeidner, Boekaerts, & Pintrich, 2000). Zimmerman and Risemberg (1997) proposed three kinds of self-regulatory strategies involved in the deliberate management of the composing process: (a) environmental strategies entail the self-regulation of the physical or social setting where writing takes place, (b) behavioral strategies comprise writing-related motoric activities, and (c) personal strategies encompass cognitive and affective processes that writers use to increase their effectiveness. Two of the most important cognitive self-regulatory strategies for organizing, producing, and transforming written text are planning and revising (Graham & Harris, 2000; Harris, Santangelo, & Graham, 2010; Zimmerman & Risemberg, 1997).

Planning involves setting goals, generating, and organizing ideas (Hayes & Flower, 1980). As it can occur before or during writing, a distinction was made between advanced and online planning (Berninger & Swanson, 1994). Several correlational studies have analyzed how students' ability to generate a plan before writing is related to their writing performance. In the studies reviewed below, preplanning skills were assessed through the complexity of students' written plans. Generally, outlines and graphic organizers are considered as the most sophisticated form of preplanning (see Hayes & Nash, 1996, for a review on planning measures).

In Grades 2 and 4, it was found that students' plans did not predict writing quality (Olinghouse & Graham, 2009). Likewise, in Grades 4–6, preplanning skills were not related to compositional quality (Whitaker, Berninger, Johnston, & Swanson, 1994). Only in Grades 7–9, positive but weak correlations were found between preplanning and writing quality ($r > .17$; Berninger, Whitaker, Feng, Swanson, & Abbott, 1996). As younger students' written plans were very similar to their texts, it was suggested that they were not differentiating planning from translating (Bereiter & Scardamalia, 1987; Berninger & Swanson, 1994; McCutchen, 2006). Moreover, it was found that only 15% of sixth graders engaged in outlining before writing (Torrance, Fidalgo, & García, 2007). This value increased to 33% in a similar study with eighth graders (Fidalgo, Torrance, & García, 2008).

Concerning revision, there is general agreement that at least it includes two key-processes: problem detection, which includes schema-guided reading and text evaluation, and problem correction, which involves the selection of a revising strategy and its implementation (Chanquoy, 2009; Fitzgerald, 1987). Whether the revising strategy operates at the surface or meaning level, it can be classified as editing or rewriting (Allal, Chanquoy, & Largy, 2004). In a similar way to preplanning, revision is hardly included in the composition process of novice writers (Fitzgerald & Markham, 1987; McCutchen, 2006). Although ability to revise emerged in Grades 4–6 in a sample studied by Whitaker et al. (1994), it only operated at all levels of language (i.e., word, sentence, and text) in Grades 7–9 (Berninger et al., 1996). Young writers' revisions seem also to have a very limited impact on text quality (Graham, Harris, MacArthur, & Schwartz, 1991)—probably because younger students tended to focus their revisions on surface problems, whereas older writers focused on meaning problems (Graham, Schwartz, & MacArthur, 1993; Harris et al., 2010; MacArthur & Graham, 1987).

Intervention studies have provided strong support for the association between planning and revision with writing quality. Meta-analyses have shown that students from Grades 2 to 12 wrote better texts after receiving instruction in planning and/or revision (Graham, McKeown, Kihara, & Harris, 2012; Graham & Perin, 2007). Importantly, writing quality increased when these strategies were taught in tandem with other self-regulatory strategies (Brunstein & Glaser, 2011; Glaser & Brunstein, 2007; for a review, see Harris & Graham, 2009). Examining the underlying mechanisms of a successful self-regulation-based intervention, Brunstein and Glaser (2011) found that it had a positive impact on text quality by promoting planning and revising. Of great import, they showed that the intervention was associated with an increase in students' writing knowledge and self-efficacy.

Writers' beliefs about their writing ability are a main component of self-regulation (Zimmerman, 1995). Self-efficacy depends on the effectiveness of the self-regulatory strategies employed and influences their persistent use in writing (Zimmerman & Risemberg, 1997). For instance, if writers attain their goals by planning or revising, their self-efficacy increases, and they continue using these strategies (Schunk & Ertmer, 2000). Consequently, writing performance is enhanced (for reviews see Klassen, 2002b; Pajares, 2003). Indeed, at different school levels, self-efficacy predicted writing quality above and beyond previous performance (effect sizes ranged from .19 to .40; Pajares, Miller, & Johnson, 1999; Pajares & Valiante, 1997, 1999). Analyzing the development of writing self-efficacy, Pajares, Valiante, and Cheong (2007) found a decrease from Grade 4 to 8. Despite the expectation that an increase in competence across schooling would be accompanied by an increase in self-efficacy, this pattern was not verified. Possibly, younger students may overestimate their writing skills, as some students with learning disabilities tend to do (Klassen, 2002a, 2002b).

Transcription Competes With Self-Regulation

Low-level transcription and high-level self-regulation processes impose heavy demands on the limited capacity of working memory. Vanderberg and Swanson (2007) showed that the central executive significantly predicted planning, translating, and revising, as well as vocabulary, punctuation, text structure, and grammar (beta weights ranged from .21 to .32). As transcription and self-regulation compete for the same pool of attentional resources, these processes must be juggled to manage cognitive load (Alamargot, Plane, Lambert, & Chesnet, 2010; Berninger, 1999; Fayol, 1999; Kellogg, 1996; McCutchen, 1996).

Beginning writers, who adopt the so-called knowledge telling strategy for composing, do not show this coordination (Bereiter & Scardamalia, 1987). Bourdin and Fayol (1994, 2000) showed that as transcription is a large resource drain, it constrains the acquisition and use of high-level writing skills (see also Alves, Branco, Castro, & Olive, 2012; Grabowski, 2010; Olive & Kellogg, 2002). This may explain, first, why young writers' barely plan or revise spontaneously and, second, why their planning and revising skills are not sufficiently developed to influence text production. However, in the course of the school years, transcription becomes more efficient, reducing the cognitive effort required (Kellogg, 2008; McCutchen, 1988; Olive, Favart, Beauvais, & Beauvais, 2009). In line with a capacity theory of writing, this gradual automatization enables writers to use their spare attentional resources for high-level processes (Fayol, 1999; McCutchen, 1996). This shift of cognitive resources allocation may set the basis for the more elaborated composing strategy of knowledge-transforming (Bereiter & Scardamalia, 1987). Transcription stops being a major source of constraint, leading to the development and successful employment of planning and revising strategies in writing.

Regarding writing self-efficacy, little is known about how it is influenced by transcription processes, which are crucial in developing writing. Given that young writers consider writing transcription features as the most important ingredients in good writing (Graham et al., 1993; Lin, Monroe, & Troia, 2007; Olinghouse & Graham, 2009), it seems likely that they may use observable

information, such as the length of their texts or the number of spelling errors, to appraise their writing ability. Indeed, one of the most influential sources of self-efficacy is students' interpretation of their own performances (Bandura, 1997).

Overview of the Current Study

Multiple-group structural equation modeling was used to examine the development of writing throughout school years. In particular, we aimed to analyze (a) the relationship between transcription (handwriting and spelling), planning, revision, self-efficacy, and the quality of text generation (story and opinion essay) and (b) if the strength of this relationship changes over time. For that, we tested the model depicted in Figure 1 at Grades 4–6 (age 9–12) and 7–9 (age 12–15). Although the proposed paths were based on the multiple sources of evidence reviewed above, to the best of our knowledge, no such model was previously tested across development.

In Grades 4–6 we predicted a direct effect of transcription on text generation quality, but in Grades 7–9 we predicted an indirect effect of transcription on text generation via planning and revision. As younger students have not mastered transcription yet, text generation was expected to be largely constrained by it (Graham et al., 1997). A different pattern was expected in older students when transcription becomes automatized and should exert less constraint on text generation (Berninger, 1999; Berninger & Swanson, 1994; Kellogg, 2008). This increased transcription fluency may enable them to develop their planning and revising abilities (Bereiter & Scardamalia, 1987; Fayol, 1999; McCutchen, 1996), which, in turn, may influence writing quality (Graham & Harris, 2000). As in Grades 7–9 (Berninger et al., 1996), but not in Grades 4–6 (Whitaker et al., 1994), planning and revising were found to be correlated, albeit weakly ($r_s = .25$), we expected a stronger effect from planning to revision in older than younger writers.

The hypotheses regarding the paths from transcription, planning, and revising to self-efficacy were as follows. In Grades 4–6, we predicted that self-efficacy would be influenced by transcription. This prediction stems not only from the critical role that transcription has on younger students' writing (Berninger, 1999) but also from their emphasis on production factors when defining good writing (Olinghouse & Graham, 2009). In Grades 7–9, we predicted that self-efficacy would be influenced by planning and revising because self-efficacy depends on the effectiveness of the self-regulatory strategies (Zimmerman & Risemberg, 1997). Older students not only use them successfully (Berninger et al., 1996) but also acknowledge their importance in writing (Graham et al., 1993). Finally, we hypothesized that self-efficacy would influence text generation at both grade levels. Research findings have shown that self-efficacy predicts writing performance throughout schooling (Pajares, 2003).

Method

Participants

Participants were 419 Portuguese native speakers in Grades 4–9. Forty-three students were excluded from the analyses based on one or more of the following criteria: absence in one of the two

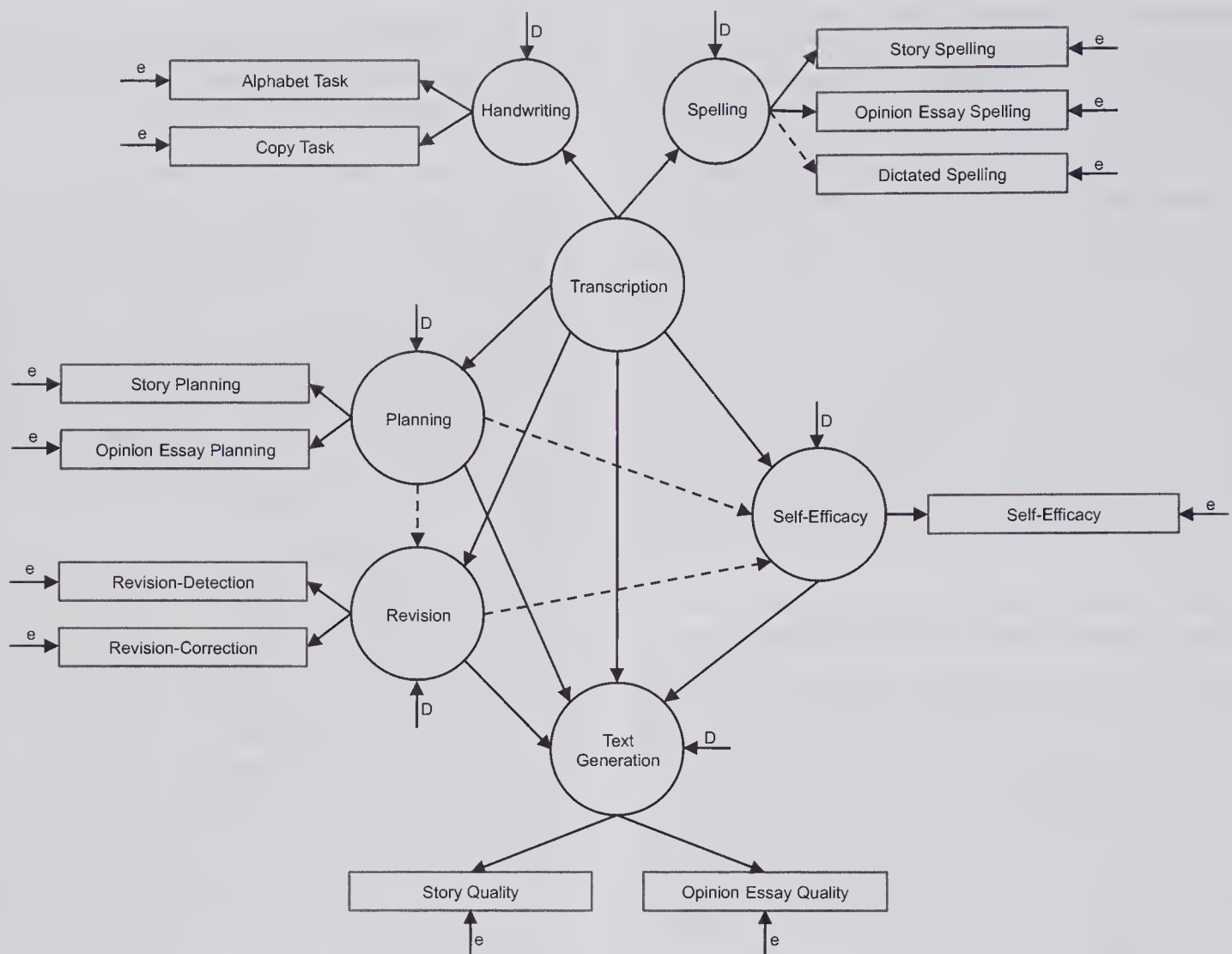


Figure 1. Structural model of the relationship between transcription, planning, revision, self-efficacy, and text generation. Circles represent factors (i.e., latent variables), rectangles represent indicators (i.e., observed variables), and arrows represent direct paths (dashed lines represent paths that were removed from the final model). e = measurement error; D = structural error.

administration sessions (17 students), task instructions not followed (22 students), special education needs (five students), and incomplete tasks (six students). Subsequent analyses were based on the data from 376 students.

Younger sample. This sample included 171 students in Grades 4–6 (57 fourth graders, $M_{\text{age}} = 10.0$ years, $SD = 0.3$, age range = 9.4–11.0; 49 fifth graders, $M_{\text{age}} = 11.0$ years, $SD = 0.6$, age range = 10.4–13.0; 65 sixth graders, $M_{\text{age}} = 12.1$ years, $SD = 0.5$, age range = 11.4–14.0; for the all sample: $M_{\text{age}} = 11.1$ years, $SD = 1.0$; 92 girls and 79 boys). Students' socioeconomic status was assessed through the educational level of their parents. Respectively, mother and father's educational level was as follows: 18% and 23% completed Grade 4 or less; 45% and 53% completed Grade 9 or less; 19% and 13% completed high school; 16% and 7% completed college or college plus some postgraduate study; and 2% and 4% was unknown. In 2011, Portuguese national statistics regarding females and males' educational level is as follows: 24% and 27% completed Grade 4 or less; 30% and 38% completed Grade 9 or less; 17% and 17% completed high school; 15% and 11% completed college or college plus some postgrad-

uate study, and 14% and 7% was unknown (Fundação Francisco Manuel dos Santos, 2012). Student's achievement was assessed via their previous marks for Portuguese, mathematics, and history. Their marks are given in a scale ranging from 1 (*lowest score*) to 5 (*highest score*). Taken all subjects together, 14% to 19% had marks below 3; 36% to 46% had marks equal 3; and 35% to 50% had marks above 3.

Older sample. This sample included 205 students in Grades 7–9 (69 seventh graders, $M_{\text{age}} = 13.0$ years, $SD = 0.4$, age range = 11.9–14.4; 61 eighth graders, $M_{\text{age}} = 13.9$ years, $SD = 0.4$, age range = 12.7–15.3; 75 ninth graders, $M_{\text{age}} = 15.0$ years, $SD = 0.5$, age range = 14.4–16.8; for the all sample, $M_{\text{age}} = 14.0$ years, $SD = 0.9$; 97 girls and 108 boys). Respectively, mother and father's educational level was as follows: 13% and 14% completed Grade 4 or less; 45% and 48% completed Grade 9 or less; 20% and 17% completed high school; 20% and 17% completed college or college plus some postgraduate study; and 2% and 4% was unknown. Regarding students' achievement, taken Portuguese, mathematics, and history together, 8% to 26% had marks below 3; 49% to 53% had marks equal 3; and 25% to 39% had marks above 3.

Setting

Students came from 19 classes integrated in a public cluster of schools located in an urban district in Northwest Portugal. In Portugal, basic education lasts 9 years and comprises three stages: Grades 1–4 (age 6–10), Grades 5–6 (age 10–12), and Grades 7–9 (age 12–15). Stage 1 is provided in primary schools and only one teacher is responsible for teaching four main courses; Stage 2 is provided in basic schools and children have one teacher for each of the nine courses; finally, Stage 3 is provided in basic or secondary schools and students have 11 courses taught by different teachers.

Regarding the teaching of writing in Portugal, two key shifts occurred in the past two decades (Álvares Pereira, Aleixo, Cardoso, & Graça, 2010). First, writing was assumed as a specific teaching object since its importance in students and professionals' lives was recognized. Second, there was a shift from a product to a process approach to writing, which provides explicit teaching on how planning, translating, and revising processes can be carried out in text production. Although writing is the preferred learning and assessment tool across courses and schooling, explicit writing instruction only occurs in Portuguese language classes.

Handwriting Fluency Measures

Alphabet task. Students were asked to write the alphabet in lowercase as quickly as possible without making mistakes (Berninger et al., 1992). The experimenter told them to stop 15 s after they had started writing the alphabet. The final score was the number of correct letters written. A letter was counted when it was legible out of context and in the right alphabetical order.

Copy task. Students were asked to copy a 60-word paragraph as quickly as possible without making mistakes. After 90 s copying it, the experimenter told them to stop. The final score was the number of words copied accurately. A word was considered correct when its letters and diacritics were clearly copied without any mistake.

Spelling Measures

Spontaneous spelling. A measure of spelling in a functional communicative context was provided by the percentage of words spelled correctly in the story and in the opinion essay.

Dictated spelling. Forty words were dictated at intervals of 6 s. These words belong to five categories representing some complexities of the Portuguese spelling system: silent letter *h*, contextual effect, position effect, inconsistency, and consonantal group (for greater detail, see Carvalhais & Castro, 2012). The final score was the total number of words spelled correctly.

Planning Measures

The experimenter gave students a green sheet and explained to them that before writing the text they would have 3 min to plan it. They were told to use that sheet as their “think pad” and to write down everything that could help them to write the text (for a similar procedure, see Berninger et al., 1996). The developmental maturity of students' planning behavior was measured with a scale ranging from 1 (*low*) to 6 (*high*). The scores 1 and 2 were attributed to plans that represent no preplanning and minimal preplanning, respectively. Plans summarizing the text received a

score of 3, and plans with topics slightly elaborated in the text received a score of 4. The scores 5 and 6 were attributed to plans with emergent subordination (i.e., rudimentary macrostructure) and structural relationships (e.g., graphic organizers), respectively. This scoring scale is non-genre-dependent and was based on the scales developed by Whitaker et al. (1994), and Olinghouse and Graham (2009). Participants made one plan for the story and another for the opinion essay and both measures were considered.

Revision Measures

To measure students' revising skills, they were asked to revise a narrative text, which had two meaning errors of three kinds created by missing, inconsistent, and out-of-sequence sentences. As younger students seem to have problems in detecting errors (Beal, 1990), which is necessary for their correction, the task was performed in two phases. First, students were asked to mark “anything that it is not right or does not sound good.” Second, the experimenter gave them the same text with the target errors marked and asked students to correct them. Respectively, the final scores were the total number of errors accurately detected (revision-detection) and corrected (revision-correction).

Self-Efficacy Measure

To measure self-efficacy beliefs, students filled out the Writing Skills Self-Efficacy scale (Pajares & Valiante, 1999) that we adapted to the Portuguese language. The scale has 10 items, which measure students' confidence about being able to accomplish specific writing skills (e.g., “Correctly spell all words in a one-page story or composition”). The answers were given in a scale ranging from 0 (*no chance*) to 100 (*completely certain*). As suggested by Pajares (2003), the self-efficacy assessment must be matched to and in close temporal proximity with the writing outcome. Accordingly, after the text topic was presented, students were asked to judge their confidence in accomplishing those skills when writing about that topic. Thus, two measures of self-efficacy were collected: story self-efficacy ($\alpha_{4-6} = .93$; $\alpha_{7-9} = .94$) and opinion essay self-efficacy ($\alpha_{4-6} = .94$; $\alpha_{7-9} = .94$). Because multicollinearity between these two measures ($r_{4-6} = .81$; $r_{7-9} = .87$) could create estimation and inference problems, as suggested by Kline (2005), they were averaged to form a composite score (viz., self-efficacy).

Text Generation Measures

Text generation was assessed through the quality of a story (“Tell a story about a child who lost his or her pet”) and an opinion essay (“Do you think teachers should give students homework every day?”). To control for potential effects of genre difficulty on subsequent tasks, writing order for genre was counterbalanced. Students had 8 min to write the text, and they were notified 4 and 2 min before the end of the time limit. Anytime a student stopped writing he or she was prompted once to continue.

Four graduate students, blind to study purposes, rated the overall text quality using a scale ranging from 1 (*low quality*) to 7 (*high quality*). To control for expected differences between grade levels, one pair of judges rated the texts from Grades 4–6, and the other pair rated the texts from Grades 7–9. Raters were told to consider

and give the same weight to the following factors: ideas quality (i.e., originality and relevance of the ideas), organization (i.e., coherence and organization of the text), sentence structure (i.e., syntactic correctness and diversity of the sentences), and vocabulary (i.e., diversity, interest, and proper use of the words). To avoid biased judgments all texts were previously typed and corrected for spelling, punctuation, and capitalization errors. For each text genre, the scores were the average for the two judges.

Reliability of Measures

At each grade level, a second judge rescored the written products for 20% of the students. For the alphabet and copy task, story and opinion essay spelling, dictated spelling, story and opinion essay planning, and error detection and correction tasks, inter-rater reliability (Pearson's coefficient) was .98, 1.00, .99, .99, 1.00, .89, .89, 1.00, and 1.00, respectively. For story and opinion essay quality evaluation, inter-rater reliability was, respectively, .79 and .84 for Grades 4–6 and .85 and .83 for Grades 7–9.

Procedure

Classroom groups with 20–25 students performed the tasks that were distributed between two 45-min sessions during the month of May (end of Portuguese academic year). Both sessions started with the presentation of the text topics. Then, students filled out the self-efficacy scale about the presented genre. After that, they planned and wrote the text. Last, students performed the spelling and revision tasks in the first session, and the copy and alphabet tasks in the second one. Two adults were always present in the room to guarantee that experimental procedures were carried out as intended.

Results

Preliminary Data Analysis

Descriptive statistics for the observed variables for Grades 4–6 and 7–9 are displayed in Table 1. The inspection of the skewness

and kurtosis of all variables revealed no distributional problems, as the absolute values of these indexes did not exceed 3.0 and 10.0, respectively (Kline, 2005). Table 2 presents the intercorrelations among all study variables by grade group. Generally, correlations were positive and modest in size, with a similar pattern for both samples.

Structural Equation Modeling

Figure 1 depicts the model that was tested against data from two groups: Grades 4–6 versus Grades 7–9. Multiple-group structural equation modeling was used to evaluate model invariance across both groups. To test the hypotheses that the relationships among latent constructs were different across samples, data analyses encompassed a series of hierarchical steps (Byrne, 2010; Kline, 2005). First, we tested if the model fit the data of both grade groups, separately. For that, single-group analyses were conducted to establish a baseline model for each group (baseline model). Second, we tested if this model fit the data of the two groups, simultaneously. For that, the parameters estimated in the baseline model were estimated in a multiple-group model, with no restrictions on its parameters (configural model). Third, we tested if the path coefficients between latent variables and indicators were equivalent. For that, factor loadings were constrained to be equal across groups (measurement model). Fourth, we examined whether factor structure was consistent across grade groups. To test structural invariance, equality constraints on structural paths were introduced in a stepwise fashion (structural model).

To evaluate fit of the models we used the chi-square statistic (χ^2), the confirmatory fit index (CFI) and the root-mean-square error of approximation (RMSEA). CFI values greater than .95 and .90, and RMSEA values less than .06 and .10 are considered good and adequate fits, respectively (Hu & Bentler, 1999). As suggested by Byrne (2010), we used the χ^2 and CFI difference tests to test for group invariance. Evidence of noninvariance is claimed when $\Delta\chi^2$ is statistically significant and Δ CFI is greater than or equal to .01 (Chen, 2007; Cheung & Rensvold, 2002).

Table 1
Descriptive Statistics for All Measures by Grade Group

Measure	Grades 4–6 (<i>n</i> = 171)				Grades 7–9 (<i>n</i> = 205)			
	<i>M</i>	<i>SD</i>	<i>Sk</i>	<i>Ku</i>	<i>M</i>	<i>SD</i>	<i>Sk</i>	<i>Ku</i>
Alphabet task	14.69	5.10	0.62	0.27	20.93	5.43	−0.02	0.49
Copy task	29.99	5.60	−0.16	−0.11	40.16	5.88	−0.39	0.44
Story spelling	95.71	4.18	−1.81	4.13	98.03	2.13	−1.70	3.39
Opinion essay spelling	95.11	4.98	−2.41	8.40	97.87	2.56	−2.43	8.19
Dictated spelling	30.71	4.44	−1.02	1.05	35.16	2.84	−1.36	2.59
Story planning	2.38	1.28	0.23	−1.64	3.10	1.39	−0.22	−1.01
Opinion essay planning	1.92	1.14	1.03	−0.25	3.06	1.39	−0.17	−1.35
Revision-detection	1.07	1.03	0.64	0.04	1.55	1.23	0.67	0.20
Revision-correction	1.32	0.94	0.23	0.14	1.75	1.03	−0.01	−0.02
Self-efficacy	73.58	17.72	−0.75	0.24	71.88	13.76	−0.34	0.10
Story quality	4.35	1.22	−0.49	0.55	3.84	1.44	−0.05	−0.34
Opinion essay quality	3.70	1.28	−0.18	−0.26	3.73	1.35	0.03	−0.36

Note. Sk = skewness; Ku = kurtosis. Metric and possible range for reported measures are as follows: alphabet task = number of correct letters; copy task = number of correct words; story and opinion essay spelling = percentage of correct words; dictated spelling = number of correct words (0–40); self-efficacy = scale ranging from 0 (*no chance*) to 100 (*completely certain*); story and opinion essay planning = scale ranging from 1 (*low*) to 6 (*high*); revision-detection = number of accurately detected errors (0–6); revision-correction = number of accurately corrected errors (0–6); story and opinion essay quality = scale ranging from 1 (*low*) to 7 (*high*).

Table 2
Correlations Between All Measures by Grade Group

Measure	1	2	3	4	5	6	7	8	9	10	11	12
1. Alphabet task	—	.51***	.32***	.28***	.27***	.13	.16	.14*	.21**	.38***	.31***	.23**
2. Copy task	.55***	—	.25***	.24***	.33***	.10	.09	.11	.21**	.35***	.34***	.26***
3. Story spelling	.26**	.16*	—	.56***	.43***	.12	.20**	.11	.18*	.32***	.20**	.25***
4. Opinion essay spelling	.22**	.23**	.66***	—	.55***	.18*	.19**	.14*	.18*	.32***	.26***	.23***
5. Dictated spelling	.36***	.29***	.62***	.55***	—	.16*	.18**	.26***	.19**	.35***	.34***	.29***
6. Story planning	.16*	.03	-.01	-.08	.16*	—	.52***	.15*	.16*	.14*	.28***	.31***
7. Opinion essay planning	.11	.06	.19*	.12	.19*	.39***	—	.13	.08	.23**	.31***	.34***
8. Revision-detection	.12	.05	.20**	.19*	.30***	.13	.14	—	.36***	.22**	.19**	.33***
9. Revision-correction	.28***	.17*	.17*	.18*	.35***	.08	.18*	.43***	—	.29***	.32***	.21**
10. Self-efficacy	.15**	.13	.34***	.26**	.40***	.11	.08	.12	.10	—	.50***	.41***
11. Story quality	.34***	.35***	.11	.16*	.27***	.08	.11	.23**	.27***	.18*	—	.44***
12. Opinion essay quality	.35***	.23**	.17*	.23**	.33***	.12	.25**	.28***	.35***	.29***	.39***	—

Note. Correlations for Grades 4–6 ($n = 171$) are below the diagonal and correlations for Grades 7–9 ($n = 205$) are above the diagonal.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Before model evaluation, latent variables were scaled by imposing unit of loading identification constraints (Kline, 2005). The unstandardized coefficients of the alphabet task, opinion essay spelling, opinion essay planning, revision-detection, self-efficacy, and opinion essay quality on the respective factors were fixed to 1.0. Only the variance of the Transcription factor was constrained to equal 1.0, so that the second-order factor loadings were freely estimated.

Baseline models. The first evaluation of the model revealed an adequate fit to the data for the younger sample, $\chi^2(43, N = 171) = 79.02$, $p = .001$, CFI = .93, RMSEA = .07, $P(\text{rmsea} \leq .05) = .09$, and a very good fit for the older sample, $\chi^2(43, N = 205) = 43.64$, $p = .44$, CFI = .99, RMSEA = .01, $P(\text{rmsea} \leq .05) = .96$. An analysis of the modification indices (MIs) revealed a problem in the model regarding the dictated spelling indicator. In Grades 4–6, MIs for the regression weights revealed two parameters with MIs greater than 6.0, which represented the cross-loadings of dictated spelling on the Revision and Text Generation factors. Because there was no strong theoretical basis to specify these additional parameters, and given that the Spelling factor already had two other indicators, we decided to remove the dictated spelling indicator. Also, to produce the most parsimonious model, the nonsignificant paths for both groups were deleted (viz., Planning \rightarrow Revision, Planning \rightarrow Self-efficacy, and Revision \rightarrow Self-Efficacy). As the effect of revision on text generation was marginally significant in both samples ($ps > .08$), we decided not to remove it. After this respecification, the final model provided a good fit to the data for Grades 4–6, $\chi^2(36, N = 171) = 52.56$, $p = .04$, CFI = .95, RMSEA = .05, $P(\text{rmsea} \leq .05) = .43$, and a very good fit to the data for Grades 7–9, $\chi^2(36, N = 205) = 29.36$, $p = .77$, CFI = 1.00, RMSEA < .001, $P(\text{rmsea} \leq .05) = .99$. Table 3 presents standardized and unstandardized regression coefficients for both samples. Although only story planning in Grades 4–6 had a marginally significant factor loading ($p = .06$), all standardized factor loadings ranged from moderate to strong ($\text{range}_{4-6} = .46-.99$; $\text{range}_{7-9} = .54-.99$), indicating that the observed variables were good indicators of the latent constructs.

Transcription, planning, revision, and self-efficacy accounted for 76% and 82% of the variance in text generation quality, respectively, in Grades 4–6 and 7–9. Considering the structural part of the model, the effects of transcription on planning ($T \rightarrow P$),

revision ($T \rightarrow R$), and self-efficacy ($T \rightarrow SE$) were significant in Grades 4–6 ($\beta_{T \rightarrow P} = .33$, $p = .006$; $\beta_{T \rightarrow R} = .57$, $p < .001$; $\beta_{T \rightarrow SE} = .39$, $p < .001$) and in Grades 7–9 ($\beta_{T \rightarrow P} = .39$, $p < .001$; $\beta_{T \rightarrow R} = .58$, $p < .001$; $\beta_{T \rightarrow SE} = .69$, $p < .001$). The effect of transcription on text generation ($T \rightarrow TG$) was significant in Grades 4–6 ($\beta_{T \rightarrow TG} = .60$, $p = .01$), but it was not in Grades 7–9 ($\beta_{T \rightarrow TG} = .26$, $p = .23$). To examine the indirect effects of transcription on text generation via planning ($T \rightarrow P \rightarrow TG$), revision ($T \rightarrow R \rightarrow TG$), and self-efficacy ($T \rightarrow SE \rightarrow TG$), we used modified Sobel tests (Sobel, 1982). The indirect effects mediated by planning and self-efficacy were significant in Grades 7–9 ($\beta_{T \rightarrow P \rightarrow TG} = .15$, Sobel $z = 2.55$, $p = .01$; $\beta_{T \rightarrow SE \rightarrow TG} = .21$, Sobel $z = 2.05$, $p = .04$), but they were not in Grades 4–6 ($\beta_{T \rightarrow P \rightarrow TG} = .03$, Sobel $z = 0.69$, $p = .49$; $\beta_{T \rightarrow SE \rightarrow TG} = .03$, Sobel $z = 0.66$, $p = .51$). The indirect effect of transcription on text generation via revision was significant in neither group ($ps > .10$). These results suggest that, for younger students, transcription contributes directly to text generation, but, for older students, transcription contributes indirectly to text generation, through planning and self-efficacy. As the baseline model was very good for both groups, invariance evaluation was conducted to analyze grade-group differences (see Table 4 for goodness-of-fit statistics).

Configural model. As the multiple-group model fitted the data very well, $\chi^2(72, N = 376) = 81.93$, $p = .20$, CFI = .99, RMSEA = .02, $P(\text{rmsea} \leq .05) = .99$, we proceeded with invariance testing.

Measurement model. The model with constrained factor loadings showed no decrement in fit, $\chi^2(77, N = 376) = 86.58$, $p < .21$, CFI = .99, RMSEA = .02, $P(\text{rmsea} \leq .05) = 1.00$, with χ^2 and CFI difference tests supporting noninvariance. Thus, there were no differences in factor loadings between Grades 4–6 and 7–9, indicating that the measures had the same meaning for both groups. After establishing measurement invariance, structural differences were examined.

Structural model. There was a decrement in fit when factor loadings and structural paths were constrained to be equal across groups, $\chi^2(86, N = 376) = 102.91$, $p < .10$, CFI = .98, RMSEA = .02, $P(\text{rmsea} \leq .05) = .99$. As the χ^2 difference test was marginally significant, and the CFI difference test supported noninvariance, we went further in the analysis to determine noninvariant paths. A stepwise procedure was used, in which only invariant

Table 3
Unstandardized and Standardized Path Coefficients by Grade Group

Path	Grades 4–6 (n = 171)		Grades 7–9 (n = 205)	
	Unstandardized	Standardized	Unstandardized	Standardized
Transcription				
Transcription → Handwriting	2.90	.67***	3.03	.76***
Alphabet task ^a	1.00	.85	1.00	.74
Copy task	0.83	.64***	1.00	.69***
Transcription → Spelling	2.15	.53***	1.23	.64***
Story spelling	0.84	.81***	0.82	.75***
Opinion essay spelling ^a	1.00	.82	1.00	.75
Planning				
Story planning	0.61	.46 <i>ns</i>	0.90***	.68***
Opinion essay planning ^a	1.00	.85	1.00	.76
Revision				
Detection ^a	1.00	.59	1.00	.54
Correction	1.15	.74***	1.03	.67***
Self-efficacy				
Self-efficacy ^b	1.00	1.00	1.00	1.00
Text generation				
Story quality	0.75	.56***	1.17	.75***
Opinion essay quality ^a	1.00	.71	1.00	.74
Transcription → Planning	0.31	.33**	0.41	.39***
Transcription → Revision	0.34	.57***	0.38	.58***
Transcription → Self-efficacy	6.83	.39***	9.38	.69***
Transcription → Text generation	0.54	.60*	0.23	.26 <i>ns</i>
Planning → Text generation	0.08	.09 <i>ns</i>	0.32	.39***
Revision → Text generation	0.44	.30 <i>ns</i>	0.33	.25 <i>ns</i>
Self-efficacy → Text generation	0.004	.09 <i>ns</i>	0.02	.31*

Note. For between-samples comparisons, see unstandardized coefficients, but for within-sample comparisons, see standardized coefficients.

^a Reference variable. ^b Single indicator of factor.

* *p* < .05. ** *p* < .01. *** *p* < .001.

paths were hold. First, we constrained the paths from transcription to handwriting and spelling. Second, we constrained the significant paths in both samples, namely, those from transcription to planning, revision, and self-efficacy. Third, we constrained the path from revision to text generation. In all of these three steps, difference tests supported noninvariance. Finally, when we constrained the paths from planning, self-efficacy, or transcription on text generation, the fit of the model declined significantly, $\Delta\chi^2(1) > 4.36$, *ps* < .05; $\Delta CFI = .01$. These analyses indicated that these three paths differed significantly between grade groups. Transcription contributed more to text generation quality in Grades 4–6,

while planning and self-efficacy contributed more to text generation quality in Grades 7–9.

Discussion

Significance of Findings

The findings of the present study are in line with the not-so-simple view of writing (Berninger & Winn, 2006) by showing that transcription and self-regulation, specifically, planning, revision, and self-efficacy are crucial for text generation in developing

Table 4
Summary of the Goodness-of-Fit Statistics for Tests of Multiple-Group Invariance

Model	χ^2	<i>df</i>	$\Delta\chi^2$	Δdf	<i>p</i>	CFI	ΔCFI
Configural model	81.93	72	—	—	—	.99	—
Measurement model	86.58	77	4.65	5	.46	.99	.00
Structural model	102.91	86	16.33	9	.06	.98	.01
H → T and S → T equal	89.85	79	3.27	2	.20	.99	.00
T → P, T → R, and T → SE equal	92.70	82	2.84	3	.42	.99	.00
R → TG equal	92.85	83	0.16	1	.69	.99	.00
T → TG equal	97.22	84	4.37	1	.04	.98	.01
P → TG equal	98.31	84	5.46	1	.02	.98	.01
SE → TG equal	98.41	84	5.56	1	.02	.98	.01

Note. CFI = comparative fit index; H = handwriting; T = transcription; S = spelling; P = planning; R = revision; SE = self-efficacy; TG = text generation.

writing. The analyses indicated that the model under test was a very good description of the data for both Grades 4–6 and 7–9. Moreover, the measurement part of the model was similar across grade groups showing that the constructs had the same meaning for both groups. Notably, we showed that these skills explained 76% and 82% of the variance in writing quality in Grades 4–6 and 7–9, respectively. Of interest, we found some differences between these two groups regarding the relationship between transcription, planning, revision, self-efficacy, and text generation.

In line with our hypothesis, transcription constrained text generation in Grades 4–6 but not in Grades 7–9. This result agrees with Berninger (1999), who showed that the explained variance in writing quality by transcription decreased from Grades 4–6 to 7–9. The direct contribution of low-level skills to writing quality in younger students might reflect a lack of automaticity in transcription (Graham et al., 1997). Because developing writers struggle with the orthographic-motor and orthographic-linguistic components of writing, these components are likely to interfere with the quality of their written texts (Berninger, 1999; Bourdin & Fayol, 1994; Olive & Kellogg, 2002). This was not the case for the older sample, in which transcription had no direct effect on writing quality. A reasonable explanation is that older students' handwriting and spelling skills were sufficiently automatized to directly constrain text generation. This is not to say that these low-level skills are no longer important. On the contrary, a main result from the present study was that transcription continued to exert its influence on writing quality after Grades 4–6 but indirectly, through its impact on planning and self-efficacy.

Consistent with our predictions, older students' transcription skills contributed indirectly to text generation via planning. Still, when we scrutinized this effect, the hypothesis was only partially confirmed because transcription contributed to text generation in Grades 7–9 as much as in Grades 4–6. Thus, in both groups, the greater the transcription fluency, the better their planning skills were. Nevertheless, while these more developed planning skills were associated to better texts in Grades 7–9, they were not in Grades 4–6. Possibly, younger students lack either sufficient planning abilities or the knowledge to appropriately use them in writing (Englert, Raphael, Fear, & Anderson, 1988; Lin et al., 2007). All in all, whereas preplanning might emerge in Grades 4–6, it only seems to be sufficiently developed to be used for the benefit of text production in Grades 7–9.

Regarding self-efficacy, we found that it was influenced by transcription not only in Grades 4–6 but also in Grades 7–9. This indicates that even older students may rely on their handwriting and spelling abilities to gauge their own sense of confidence. Nonetheless, while self-efficacy influenced older students' writing quality, it did not in the younger sample. It is possible that young writers were not able to translate their perceived self-efficacy into corresponding performance. Students might have lacked the necessary knowledge and skills to proactively adjust their writing behavior to their appraisals of personal capabilities (Bandura, 1997). Although this explanation assumes that students' self-efficacy judgments were accurate, this could have not been the case. Indeed, given that self-efficacy influence task choice, expended effort, perseverance, and emotional reactions, faulty self-judgments could also explain why novice writers' writing performance was unrelated to self-efficacy.

Of concern were the results about revision, which were similar across grade groups. Although students' transcription fluency predicted students' skills to revise meaning errors, these skills were not related to writing quality. This latter result might be explained differently according to grade group. It is possible that younger students lacked sufficient revising skills. By contrast, it might be that older students, albeit being in the possession of those skills, did not use them to increase the quality of their writing. It could be argued that students did not have enough time to employ their revising skills in an 8-min writing task. This was probably not the case because, in a writing task without time limits, eighth graders only spent 10% of their writing time revising their texts (Fidalgo et al., 2008). As revision places large demands on working memory, it is possible that older students were not able to write their texts and, simultaneously, revise them for meaning (Hacker, 1994). Probably, postponing revision would have improved text quality (Chanquoy, 2001).

Finally, the predicted relationship between the self-regulation variables in Grades 7–9 was not found. In the sample studied, writers' ability to generate written plans before writing was not linked to their ability to revise meaning errors, suggesting that these skills did not develop in tandem. This result might be explained by the different nature of these strategies: Writers plan what they are going to write, but they revise what they have already written. In addition, the lack of relationship between planning and revising is possibly related to the finding that while some students tend to adopt planning strategies, others tend to prefer revising strategies (Kieft, Rijlaarsdam, Galbraith, & van der Bergh, 2007). Unexpectedly, the paths from planning and revising to self-efficacy were also nonsignificant. This result might be related to the use of a general self-efficacy measure, not explicitly tied to the use of writing self-regulatory strategies. Bruning, Dempsey, Kauffman, McKim, and Zumbrunn (2012) found empirical support for a three-factor model of writing self-efficacy comprising self-efficacy for writing ideation, writing conventions, and writing self-regulation. The assessment of specific dimensions of self-efficacy, such as *self-efficacy for self-regulated learning* (Zimmerman & Martinez-Pons, 1990), can inform us better about how students' beliefs are influenced by their planning and revising skills.

Limitations and Future Research Directions

Some limitations in the present study need to be considered, as well as possible ways to further explore the development of writing. First, the data came from a single group of schools. However, the sample included a full range of backgrounds and the main results confirmed the literature reviewed.

Second, by asking students' to plan and revise, we do not know if they were able to do it spontaneously in their texts. Indeed, it is as important to have the appropriate skills to use a strategy, as to autonomously decide when to employ that strategy. Future research should therefore focus on the extent to which students can deliberately plan and revise and how this impacts writing performance.

A third limitation, which is related to the previous one, is that online planning and online revision were not examined. By analyzing the online management of these processes we could deepen

our understanding about their interaction and temporal distribution as a function of transcription.

Fourth, working memory and writing knowledge were not included in the model. Working memory is a pivotal system in the relationship between low- and high-level writing processes (Kellogg, 1996; McCutchen, 1996). The inclusion of a working memory factor could have provided valuable information about the evolution of this relationship during school years. Also, the students' writing knowledge and its impact on writing has been widely discussed in the literature (Englert et al., 1988; Graham et al., 1993; Lin et al., 2007; McCutchen, 2011). Very early on, knowledge about writing predicted writing quality, above and beyond transcription and self-regulation (Olinghouse & Graham, 2009). The relationship of writing knowledge with these processes deserves further attention.

Finally, any conclusion drawn from our results is limited to the indicators used and to writing assessment, as writing instruction was not studied in this project. Additional self-regulatory strategies, such as goal-setting, self-monitoring, or self-instructions (Graham & Harris, 2000; Harris et al., 2010; Zimmerman & Risemberg, 1997), should be examined. Likewise, as intraindividual differences at the text, sentence, and word levels were found (Wagner et al., 2011; Whitaker et al., 1994), other text generation measures should be considered in future research.

Educational Implications

This study confirmed that transcription contributes to developing writing (Berninger & Swanson, 1994; Graham et al., 1997) and is likely to hamper the acquisition and development of high-level writing processes, which characterizes mature writing (Alamargot et al., 2010). For that reason, transcription should be taught and practiced until a proficient level of automaticity is achieved. Indeed, through its influence on planning maturity and self-efficacy beliefs, transcription stills constraining older students' writing. Educational research has already shown the positive effects of interventions targeting handwriting (e.g., Christensen, 2004; Jones & Christensen, 1999) and spelling (e.g., Berninger et al., 1998, 2002; Graham, Harris, & Fink-Chorzempa, 2002). In spite of that, these skills tend to be neglected by teachers beyond the initial years of learning to write.

The findings that in Grades 4–6 self-regulation variables were influenced by transcription but did not influence text quality, suggest that this developmental age may be a sensitive period to promote planning and revising as well as to nurture self-efficacy beliefs. Particular attention should be given to the development of revising skills because even older students do not seem to use them as an aid to write better texts. It has been widely demonstrated that teaching self-regulatory strategies builds self-efficacy and enhances writing quality (see Harris & Graham, 2009, for further discussion). Even though it is not desirable that these skills become fully automatized (McCutchen, 1988), through teaching, they can become fluent and increase writing efficiency. To fulfill students writing needs, the design of intervention programs tapping low- and high-level skills is clearly warranted (for successful programs see Berninger et al., 2006; Berninger et al., 2002).

In conclusion, the present study analyzed the role of transcription and self-regulation in text generation quality throughout development. Transcription proved to be the most restrictive factor to

writing quality, directly, in Grades 4–6, and, indirectly via planning and self-efficacy, in Grades 7–9. Our study adds to a growing body of research showing that writing development is heavily based on transcription and self-regulation. If we want to enhance students' written composition across school years, none of these sets of skills should be left behind.

References

- Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85*, 478–508. doi:10.1037/0022-0663.85.3.478
- Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in Grades 1 to 7. *Journal of Educational Psychology, 102*, 281–298. doi:10.1037/a0019318
- Alamargot, D., Plane, S., Lambert, E., & Chesnet, D. (2010). Using eye and pen movements to trace the development of writing expertise: Case studies of a 7th, 9th and 12th grader, graduate student, and professional writer. *Reading and Writing, 23*, 853–888. doi:10.1007/s11145-009-9191-9
- Allal, L., Chanquoy, L., & Largy, P. (Eds.). (2004). *Revision: Cognitive and instructional processes*. Dordrecht, the Netherlands: Kluwer Academic. doi:10.1007/978-94-007-1048-1
- Álvares Pereira, L., Aleixo, C., Cardoso, I., & Graça, L. (2010). The teaching and learning of writing in Portugal: The case of a research group. In C. Bazerman, R. Krut, K. Lunsford, S. Mcleod, S. Null, P. Rogers, & A. Stansell (Eds.), *Traditions of writing research* (pp. 58–70). New York, NY: Routledge.
- Alves, R. A., Branco, M., Castro, S. L., & Olive, T. (2012). Effects of handwriting skill, handwriting and dictation modes, and gender of fourth graders on pauses, written language bursts, fluency, and quality. In V. W. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 389–402). New York, NY: Psychology Press.
- Alves, R. A., & Jesus, I. (2011). *Bursts of written language production increase across the initial years of schooling*. Paper presented at the 14th biennial conference of the European Association for Research on Learning and Instruction, Exeter, England.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.
- Beal, C. R. (1990). The development of text evaluation and revision skills. *Child Development, 61*, 247–258. doi:10.2307/1131063
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.
- Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatic and constructive processes. *Learning Disability Quarterly, 22*, 99–112. doi:10.2307/1511269
- Berninger, V. W., & Chanquoy, L. (2012). What writing is and how it changes across early and middle childhood development: A multidisciplinary perspective. In E. L. Grigorenko, E. Mambrino & D. D. Preiss (Eds.), *A mosaic of new perspectives* (pp. 65–84). London, England: Psychology Press.
- Berninger, V. W., Rutberg, J. E., Abbott, R. D., Garcia, N., Anderson-Youngstom, M., Brooks, A., & Fulton, C. (2006). Tier 1 and tier 2 early intervention for handwriting and composing. *Journal of School Psychology, 44*, 3–30. doi:10.1016/j.jsp.2005.12.003
- Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In E. C. Butterfield (Ed.), *Children's writing: Toward a process theory of the development of skilled writing* (Vol. 2, pp. 57–81). Greenwich, CT: JAI Press.

- Berninger, V. W., Vaughan, K. B., Abbott, R. D., Begay, K., Coleman, K. B., Curtin, G., . . . Graham, S. (2002). Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology, 94*, 291–304. doi:10.1037/0022-0663.94.2.291
- Berninger, V. W., Vaughan, K. B., Abbott, R. D., Brooks, A., Abbott, S. P., Rogan, L. W., . . . Graham, S. (1998). Early intervention for spelling problems: Teaching functional spelling units of varying size with a multiple-connections framework. *Journal of Educational Psychology, 90*, 587–605. doi:10.1037/0022-0663.90.4.587
- Berninger, V. W., Whitaker, D., Feng, Y., Swanson, H. L., & Abbott, R. D. (1996). Assessment of planning, translating, and revising in junior high writers. *Journal of School Psychology, 34*, 23–52. doi:10.1016/0022-4405(95)00024-0
- Berninger, V. W., & Winn, W. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). New York, NY: Guilford Press.
- Berninger, V. W., Yates, C. M., Cartwright, A. C., Rutberg, J., Remy, E., & Abbott, R. D. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing, 4*, 257–280. doi:10.1007/BF01027151
- Bourdin, B., & Fayol, M. (1994). Is written language production more difficult than oral language production? A working memory approach. *International Journal of Psychology, 29*, 591–620. doi:10.1080/00207599408248175
- Bourdin, B., & Fayol, M. (2000). Is graphic activity cognitively costly? A developmental approach. *Reading and Writing, 13*, 183–196. doi:10.1023/A:1026458102685
- Bruning, R. H., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2012, August 13). Examining dimensions of self-efficacy for writing. *Journal of Educational Psychology*. Advance online publication. doi:10.1037/a0029692
- Brunstein, J. C., & Glaser, C. (2011). Testing path-analytic mediation model of how self-regulated writing strategies improve fourth graders' composition skills: A randomized controlled trial. *Journal of Educational Psychology*. Advance online publication. doi:10.1037/a0024622
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.
- Carvalho, L., & Castro, S. L. (2012). *Lexicality, complexity, and lengths effects on the development of Portuguese spelling*. Manuscript in preparation.
- Chanquoy, L. (2001). How to make it easier for children to revise their writing: A study of revision from 3rd to 5th grades. *British Journal of Educational Psychology, 71*, 15–41. doi:10.1348/000709901158370
- Chanquoy, L. (2009). Revision processes. In R. Beard, D. Myhill, J. Riley, & M. Nystrand (Eds.), *The Sage handbook of writing development* (pp. 80–97). London, England: Sage. doi:10.4135/9780857021069.n6
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. doi:10.1080/10705510701301834
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. doi:10.1207/S15328007SEM0902_5
- Christensen, C. A. (2004). Relationship between orthographic-motor integration and computer use for the production of creative and well-structured written text. *British Journal of Educational Psychology, 74*, 551–564. doi:10.1348/0007099042376373
- Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology, 77*, 479–492. doi:10.1348/000709906X116768
- De La Paz, S., & Graham, S. (1995). Dictation: Applications to writing for students with learning disabilities. In T. Scruggs & M. Mastropieri (Eds.), *Advances in learning and behavioral disorders* (Vol. 9, pp. 227–247). Greenwich, CT: JAI Press.
- Englert, C. S., Raphael, T. E., Fear, K. L., & Anderson, L. M. (1988). Students' metacognitive knowledge about how to write informational texts. *Learning Disability Quarterly, 11*, 18–46. doi:10.2307/1511035
- Fayol, M. (1999). From on-line management problems to strategies in written composition. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing: Processing capacity and working memory effects in text production* (pp. 13–23). Amsterdam, the Netherlands: Amsterdam University Press.
- Fidalgo, R., Torrance, M., & García, J.-N. (2008). The long-term effects of strategy-focussed writing instruction for grade six students. *Contemporary Educational Psychology, 33*, 672–693. doi:10.1016/j.cedpsych.2007.09.001
- Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*, 481–506. doi:10.3102/00346543057004481
- Fitzgerald, J., & Markham, L. R. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4*, 3–24. doi:10.1207/s1532690xci0401_1
- Fundação Francisco Manuel dos Santos. (2012). *PORDATA, Base de dados Portugal contemporâneo* [PORDATA, Contemporary Portugal Database]. Retrieved from <http://www.pordata.pt/>
- Glaser, C., & Brunstein, J. C. (2007). Improving fourth-grade students' composition skills: Effects of strategy instruction and self-regulation procedures. *Journal of Educational Psychology, 99*, 297–310. doi:10.1037/0022-0663.99.2.297
- Grabowski, J. (2010). Speaking, writing, and memory span in children: Output modality affects cognitive performance. *International Journal of Psychology, 45*, 28–39. doi:10.1080/00207590902914051
- Graham, S. (1990). The role of production factors in learning disabled students' compositions. *Journal of Educational Psychology, 82*, 781–791. doi:10.1037/0022-0663.82.4.781
- Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology, 89*, 170–182. doi:10.1037/0022-0663.89.1.170
- Graham, S., & Harris, K. R. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist, 35*, 3–12. doi:10.1207/S15326985EP3501_2
- Graham, S., Harris, K. R., & Fink-Chorzempa, B. (2002). Contribution of spelling instruction to the spelling, writing and reading of poor spellers. *Journal of Educational Psychology, 94*, 669–686. doi:10.1037/0022-0663.94.4.669
- Graham, S., Harris, K. R., MacArthur, C. A., & Schwartz, S. S. (1991). Writing and writing instruction for students with learning disabilities: Review of a research program. *Learning Disability Quarterly, 14*, 89–114. doi:10.2307/1510517
- Graham, S., McKeown, D., Kiuahara, S., & Harris, K. R. (2012, July 9). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*. Advance online publication. doi:10.1037/a0029185
- Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445–476. doi:10.1037/0022-0663.99.3.445
- Graham, S., Schwartz, S. S., & MacArthur, C. A. (1993). Knowledge of writing and the composition process, attitude toward writing, and self-efficacy for students with and without learning disabilities. *Journal of Learning Disabilities, 26*, 237–249. doi:10.1177/002221949302600404
- Hacker, D. J. (1994). Comprehension monitoring as a writing process. In E. C. Butterfield (Ed.), *Children's writing: Toward a process theory of the development of skilled writing* (Vol. 2, pp. 143–172). Greenwich, CT: JAI Press.

- Harris, K. R., & Graham, S. (2009). Self-regulated strategy development in writing: Premises, evolution, and the future. *British Journal of Educational Psychology Monograph Series II, 1* (No. 6), 113–135. doi:10.1348/978185409X422542
- Harris, K. R., Santangelo, T., & Graham, S. (2010). Metacognition and strategies instruction in writing. In H. S. Waters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 226–256). New York, NY: Guilford Press.
- Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Erlbaum.
- Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–29). Hillsdale, NJ: Erlbaum.
- Hayes, J. R., & Nash, J. G. (1996). On the nature of planning in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 29–55). Mahwah, NJ: Erlbaum.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi:10.1080/10705519909540118
- Jones, D., & Christensen, C. A. (1999). Relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology, 91*, 44–49. doi:10.1037/0022-0663.91.1.44
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437–447. doi:10.1037/0022-0663.80.4.437
- Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78*, 243–255. doi:10.1037/0022-0663.78.4.243
- Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 57–71). Mahwah, NJ: Erlbaum.
- Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1*, 1–26.
- Kieft, M., Rijlaarsdam, G., Galbraith, D., & van der Bergh, H. (2007). The effects of adapting a writing course to students' writing strategies. *British Journal of Educational Psychology, 77*, 565–578. doi:10.1348/096317906X120231
- Klassen, R. (2002a). A question of calibration: A review of the self-efficacy beliefs of students with learning disabilities. *Learning Disability Quarterly, 25*, 88–102. doi:10.2307/1511276
- Klassen, R. (2002b). Writing in early adolescence: A review of the role of self-efficacy beliefs. *Educational Psychology Review, 14*, 173–203. doi:10.1023/A:1014626805572
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Lin, S.-J. C., Monroe, B. W., & Troia, G. A. (2007). Development of writing knowledge in Grades 2–8: A comparison of typically developing writers and their struggling peers. *Reading & Writing Quarterly, 23*, 207–230. doi:10.1080/10573560701277542
- MacArthur, C. A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *The Journal of Special Education, 21*, 22–42. doi:10.1177/002246698702100304
- McCutchen, D. (1988). "Functional automaticity" in children's writing: A problem of metacognitive control. *Written Communication, 5*, 306–324. doi:10.1177/0741088388005003003
- McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*, 299–325. doi:10.1007/BF01464076
- McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115–130). New York, NY: Guilford Press.
- McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research, 3*, 51–68.
- Medwell, J., & Wray, D. (2008). Handwriting: A forgotten language skill? *Language and Education, 22*, 34–47. doi:10.2167/le722.0
- Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101*, 37–50. doi:10.1037/a0013462
- Olive, T., Favart, M., Beauvais, C., & Beauvais, L. (2009). Children's cognitive effort and fluency in writing: Effects of genre and of handwriting automatisations. *Learning and Instruction, 19*, 299–308. doi:10.1016/j.learninstruc.2008.05.005
- Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory & Cognition, 30*, 594–600. doi:10.3758/BF03194960
- Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly, 19*, 139–158. doi:10.1080/10573560308222
- Pajares, F., Miller, M. D., & Johnson, M. J. (1999). Gender differences in writing self-beliefs of elementary school students. *Journal of Educational Psychology, 91*, 50–61. doi:10.1037/0022-0663.91.1.50
- Pajares, F., & Valiante, G. (1997). Influence of self-efficacy on elementary students' writing. *The Journal of Educational Research, 90*, 353–360. doi:10.1080/00220671.1997.10544593
- Pajares, F., & Valiante, G. (1999). Grade level and gender differences in the writing self-beliefs of middle school students. *Contemporary Educational Psychology, 24*, 390–405. doi:10.1006/ceps.1998.0995
- Pajares, F., Valiante, G., & Cheong, Y. F. (2007). Writing self-efficacy and its relation to gender, writing motivation and writing competence: A developmental perspective. In S. Hidi & P. Boscolo (Eds.), *Writing and motivation* (pp. 141–159). Amsterdam, the Netherlands: Elsevier.
- Reece, J., & Cumming, G. (1996). Evaluating speech-based composition methods: Planning, dictation, and the listening word processor. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 361–380). Mahwah, NJ: Erlbaum.
- Scardamalia, M., Bereiter, C., & Goleman, H. (1982). The role of production factors in writing ability. In M. Nystrand (Ed.), *What writers know: The language, process, and structure of written discourse* (pp. 173–210). New York, NY: Academic Press.
- Schunk, D. H., & Ertmer, P. A. (2000). Self-regulation and academic learning: Self-efficacy enhancing interventions. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Self-regulation: Theory, research, and applications* (pp. 631–649). Orlando, FL: Academic Press.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). Washington, DC: American Sociological Association. doi:10.2307/270723
- Torrance, M., Fidalgo, R., & García, J.-N. (2007). The teachability and effectiveness of cognitive self-regulation in sixth-grade writers. *Learning and Instruction, 17*, 265–285. doi:10.1016/j.learninstruc.2007.02.003
- Vanderberg, R., & Swanson, H. L. (2007). Which components of working memory are important in the writing process? *Reading and Writing, 20*, 721–752. doi:10.1007/s11145-006-9046-6
- Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Wilson, L. G., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing, 24*, 203–220. doi:10.1007/s11145-010-9266-7
- Whitaker, D., Berninger, V. W., Johnston, J., & Swanson, H. L. (1994). Intraindividual differences in levels of language in intermediate grade

- writers: Implications for the translating process. *Learning and Individual Differences*, 6, 107–130. doi:10.1016/1041-6080(94)90016-7
- Zeidner, M., Boekaerts, M., & Pintrich, P. R. (2000). Self-regulation: Directions and challenges for future research. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Self-regulation: Theory, research, and applications* (pp. 749–768). Orlando, FL: Academic Press.
- Zimmerman, B. J. (1995). Self-regulation involves more than metacognition: A social cognitive perspective. *Educational Psychologist*, 30, 217–221. doi:10.1207/s15326985ep3004_8
- Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, 82, 51–59. doi:10.1037/0022-0663.82.1.51
- Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology*, 22, 73–101. doi:10.1006/ceps.1997.0919

Received January 31, 2012

Revision received November 26, 2012

Accepted December 4, 2012 ■

Do Early Literacy Skills in Children's First Language Promote Development of Skills in Their Second Language? An Experimental Evaluation of Transfer

J. Marc Goodrich and Christopher J. Lonigan
Florida State University

JoAnn M. Farver
University of Southern California

The purpose of this study was to evaluate the cross-language transfer of the emergent literacy skills of preschoolers who were Spanish-speaking language-minority children in the context of an experimental intervention study. Ninety-four children were randomly assigned either to a control condition (HighScope Preschool Curriculum) or to receive small-group pull-out instruction (Literacy Express Preschool Curriculum) in English or initially in Spanish and transitioning to English. We examined whether children's initial skills in one language moderated the impact of the intervention on those same skills in the other language at posttest. Results demonstrated that for children in the English-only intervention condition, initial Spanish receptive vocabulary and elision skills moderated the impact of the intervention on English receptive vocabulary and elision skills at posttest, respectively. For children in the transitional intervention condition, initial English definitional vocabulary and elision skills moderated the impact of the intervention on Spanish definitional vocabulary and elision skills at posttest, respectively. Results for the vocabulary interactions supported the notion of transfer of specific linguistic information across languages, whereas results for the elision interaction for the English-only intervention group comparisons supported language-independent transfer. Results for the elision interaction for the transitional intervention group comparisons supported both language-independent and language-specific transfer. Implications for the theory of cross-language transfer of emergent literacy skills are discussed.

Keywords: English language learners, language minority, emergent literacy, transfer

Latinos are the largest and fastest growing segment of the U.S. population. As of 2011, the U.S. Census Bureau reported that there were more than 49 million people of Latino origin living in the United States, accounting for 16% of the population. The Latino population grew an estimated 3.2% from 2007 to 2008 (approximately 1.5 million people), and it continues to grow rapidly due to immigration from many Latin American countries. In the United States, 26% of the population age 5 years or younger as well as 23% of the population age 18 years or younger are of Latino descent. Latino students now represent the second largest population of students within the United States (Hemphill, Vanneman, &

Rahman, 2011). Children who are exposed to a significant amount of Spanish in the home make up a large portion of the Latino population in the United States and are at a particularly high risk for developing reading problems. The U.S. Census Bureau (2007) reported that 12.3% of the U.S. population older than age 5 speaks Spanish or a Spanish Creole at home. Latino children who speak Spanish are often referred to as either Spanish-speaking English language learners (ELLs) or language-minority (LM) children. An important distinction between these two classifications is that children identified as Spanish-speaking ELLs must have limited English proficiency; however, LM refers to children who are exposed to a language other than English in the home but who do not necessarily have limited English proficiency. Therefore, LM children encompass all children who are exposed to Spanish in the home whether or not they have limited English proficiency.

Children's language background is an important factor for evaluating their risk status for reading difficulties, and U.S. Latino children are at a high risk of developing reading problems. According to Hemphill et al. (2011), there is a large gap between the reading performance of Latino children and that of White children. Although the overall reading scores of both fourth- and eighth-grade Latino students have improved from 1992 to 2009, the gap between Latino and White students has remained constant. Latino children represent a significant challenge to educators who are charged with the task of helping these children develop their reading skills and to narrow the existing performance gap between Latino and White students.

This article was published Online First February 18, 2013.

J. Marc Goodrich and Christopher J. Lonigan, Department of Psychology and the Florida Center for Reading Research, Florida State University; JoAnn M. Farver, Department of Psychology, University of Southern California.

This research was supported by National Science Foundation Grant REC-0128970. Preparation of this work was supported by grants from the National Institute of Child Health and Human Development (HD060292) and the Institute of Education Sciences (R305B090021). The views expressed herein are those of the authors and have not been reviewed or approved by the granting agencies.

Correspondence concerning this article should be addressed to J. Marc Goodrich or to Christopher J. Lonigan, Department of Psychology, Florida State University, 1107 West Call Street, Tallahassee, FL 32306. E-mail: marcgoodrich5@gmail.com or lonigan@psy.fsu.edu

Research indicates that emergent literacy skills are associated with children's later reading skills and are measurable as early as the preschool years (Whitehurst & Lonigan, 1998). Specifically, the three skills that are the most predictive of future reading ability in monolingual English-speaking children are phonological awareness (PA), print knowledge, and oral language (Lonigan, Schatschneider, & Westberg, 2008; Whitehurst & Lonigan, 1998). PA is a child's ability to detect and manipulate the sounds of spoken language independent of the semantic properties of those sounds. Print knowledge is children's understanding of how print is organized, as well as letter-name and letter-sound knowledge. Oral language consists of a child's vocabulary, as well as his or her ability to use vocabulary within context to convey and understand meaning. Understanding these precursors to reading is important because reading ability becomes relatively stable as early as kindergarten (Wagner, Torgesen, & Rashotte, 1994; Wagner et al., 1997). Although these emergent literacy skills are often correlated with one another, they are each distinct skills that uniquely predict children's later reading abilities (Whitehurst & Lonigan, 1998).

Although there is a large body of research examining reading and reading-related skills of monolingual English-speaking children, substantially less research has been conducted concerning reading and reading-related skills of LM children. Most research that has examined reading and reading-related skills of LM children indicates that most of the same factors that contribute to reading difficulty or success among monolingual English-speaking students also contribute to later reading difficulty or success among LM children (e.g., Lindsey, Manis, & Bailey, 2003; Manis, Lindsey, & Bailey, 2004). For example, word identification skills have similar developmental trajectories for both LM and monolingual English-speaking children from kindergarten through second grade (Manis et al., 2004). Similarly, skills such as phonological awareness, oral language, and print knowledge that predict reading outcomes in monolingual English-speaking students also predict reading outcomes among LM children (Lindsey et al., 2003; Manis et al., 2004).

Researchers have investigated the relations between first-language (L1) and second-language (L2) emergent literacy skills of Latino LM children, examining whether the level of proficiency of emergent literacy skills in children's L1 predicts their competency in L2 (e.g., Leafstedt & Gerber, 2005; López & Greenfield, 2004; Tabors, Pérez, & López, 2003). Research has demonstrated that for LM children, skills within the domain of PA are related both within and across languages (e.g., Branum-Martin et al., 2006; Gottardo, 2002; Gottardo & Mueller, 2009; Leafstedt & Gerber, 2005). Specifically, children with strong L1 PA skills tend to have strong L2 PA skills (e.g., Atwill, Blanchard, Gorin, & Burstein, 2007). Although some researchers claim that evidence indicating that children's L1 skills predict their competency in L2 demonstrates that children "transfer" these skills across languages, other researchers have demonstrated that L1 and L2 PA skills are separate but related constructs; that is, they form distinct factors—even though the factors are correlated with one another—when examined with confirmatory factor analysis (Branum-Martin et al., 2006; Gottardo & Mueller, 2009).

Most prior studies on the relations between L1 and L2 reading-related skills have focused on LM children in the early elementary school years. Less research has been conducted with LM children in preschool. Studies indicate that there are positive correlations

between the L1 and L2 PA skills of LM preschool children (Anthony et al., 2009; Dickinson, McCabe, Clark-Chiarelli, & Wolf, 2004; López & Greenfield, 2004; Tabors et al., 2003). More research examining cross-language relations of emergent literacy skills with LM preschool children is needed to allow researchers to determine fully if, how, and when these emergent literacy skills transfer from one language to another.

Cummins (1979) introduced the developmental interdependence hypothesis (DIH) of cross-language transfer as an attempt to explain the development of language and literacy skills of LM children. He proposed that among LM children, development of language-related skills in children's L2 is dependent upon the children's proficiency in those skills in their L1. More specifically, according to the DIH, the ability to acquire proficiency in L2 depends on the competence of the individual in L1 skills at the time of initial exposure to L2; however, this transfer is not automatic. According to the DIH, children's skills in one language will transfer to a second language only if there is sufficient exposure to that language (a characteristic of L2 input) and motivation to learn it (an attribute of the individual learning the L2; Cummins, 1981). The importance of exposure to L2 is highlighted by consistent findings that the length of residency in a country in which the primary language is the individual's L2 is strongly related to L2 acquisition (Cummins, 1991); however, attributes of the individual learning the L2 are also important for the development of strong L2 skills, as evidenced by the finding that cognitive and personality characteristics contribute as much to the development of L2 academic proficiency as does length of residency (Cummins et al., 1984).

Cummins (1981) integrated the notion of a common underlying proficiency (CUP) within the DIH as an additional perspective on the phenomenon of transfer. He stated that in addition to L1 or L2 instruction leading to the development of skills in that particular language, "experience with either language can promote development of the proficiency underlying both languages" (Cummins, 1981, p. 25). Once this CUP is developed, children are able to apply this knowledge to any subsequently learned language (i.e., to "transfer" skills from one language to another). Reviews of research evaluating the DIH have indicated that there is strong support for this theory (Fitzgerald, 1995). The notion of a CUP is similar to hypotheses advanced by others. For example, the general abilities model (e.g., Castilla, Restrepo, & Perez-Leroux, 2009) posits that the strong relation between skills in L1 and L2 represents an underlying language-learning capacity that children have independent of their intelligence or overall cognitive abilities.

In Cummins' theory (e.g., 1981, 2008), DIH and CUP are viewed as essentially the same phenomenon; however, the theory allows both for cross-linguistic relations of skills that are due to actual transfer from one language to another and for cross-linguistic relations of skills that are due to language-independent attributes of the individual that are related to performance in both languages. Understanding which process underlies cross-linguistic relations for specific skills is important to advancing knowledge concerning the development of academic skills in LM children and may have implications for assessment, identification, and instruction (Castilla et al., 2009). Therefore, in this study, we use the term CUP to refer to language-independent processes that are related to transfer and the term DIH to refer to direct transfer of specific linguistic information across languages.

Cummins (1991) proposed that two specific predictions can be drawn from the DIH: (a) L1 and L2 skills that are related across languages can be attributed to both underlying attributes of the individual who is learning the L2 as well as the quality and quantity of L2 input received, and (b) L1 and L2 skills that are not related across languages solely represent the quality and quantity of L2 input received, not underlying attributes of the individual. To the extent to which there is a large degree of overlap of sounds across languages, PA would be a specific-language-independent skill. LM children who can detect and manipulate sounds in one of their languages should also be able to detect and manipulate those same sounds in their other language. Therefore, cross-language relations between skills that are specific language independent such as PA should be related to attributes of the individual (i.e., an underlying language-learning capacity and development of a CUP) and the quality and quantity of input. In contrast, cross-language relations among skills that are language specific, such as vocabulary knowledge, should be related solely to the quality and quantity of input received.

Although results of a number of studies seem to support the DIH and the presence of a CUP (e.g., López & Greenfield, 2004; Tabors et al., 2003), all of these studies are correlational, leaving their results open to alternative interpretations. The term *transfer* implies something more than the simple co-occurrence of skills; namely, if children have received instruction on particular skills in one language, they should demonstrate gains in those same skills in their other language, provided they receive adequate exposure to their other language, either at home, with their peers, or at school. To test adequately whether transfer occurs, either experimental evidence or longitudinal data are necessary.

Farver, Lonigan, and Eppe (2009) reported the results of a study in which the impacts of two variations of a small-group early literacy intervention were evaluated relative to a business-as-usual control with Spanish-speaking LM preschool children. In one intervention condition, children received all instruction in English. In the other condition, children initially received instruction in Spanish, and instruction was transitioned to English over the preschool year. Farver et al. reported that children in both interventions ended the preschool year with significantly better scores on all measured early literacy skills than did children in the control condition. Our goal in the current study was to expand on the analysis conducted by Farver et al. to determine if the emergent literacy skills of preschool LM children transfer from one language to another in the context of instruction aimed at improving these skills. To do so, we examined whether children's initial L1 skills moderated the impact of the intervention designed to improve those skills in children's L2 and vice versa, using the same data used by Farver et al. Whereas Farver et al. evaluated the overall impact of the intervention, in this study we evaluated whether children's pretest skills moderated the impact of the intervention. This analysis represents a better test of transfer than do simple correlational studies because it examines the impact of skills in children's L1 on skills in their L2 in the context of experimentally manipulated instruction designed to improve their L2 skills. For example, if children with higher initial L1 skills benefit more from the intervention than do children with lower initial L1 skills, it could be concluded that a part of the positive impact of the intervention on L2 skills was the result of strong initial L1 skills.

To test how children's initial L1 skills moderated the impact of the intervention, we tested moderation in two steps. First, we tested the moderation of the impact of the intervention by initial L1 or L2 skills for L2 and L1 outcomes, respectively, to determine if children with greater initial skills in one language benefitted more from an intervention on those same skills in their other language. Then, the moderation of the impact of the intervention by initial skills in the same language as the outcome was added to the models. This second interaction term evaluated the degree to which the CUP across languages accounted for transfer. To demonstrate support for the transfer of specific linguistic information (i.e., a CUP-independent DIH), the initial moderation effect would have to remain significant when tested in the context of the second interaction term. In contrast, support for the CUP model would be obtained if a significant initial moderation effect were rendered nonsignificant by the inclusion of the second interaction term. Support could be obtained for both the CUP-independent DIH and the CUP model if both interaction terms were significant. For example, it is possible that children transfer specific linguistic information about skills across languages as well as utilize a CUP to benefit from instruction. Because prior research suggests that "backward" transfer (i.e., transfer from L2 to L1) can occur (e.g., Dressler & Kamil, 2006), analyses of the influence of L2 on L1 were included as well (i.e., do Spanish-speaking LM children with greater initial English skills benefit more from an intervention designed to improve their Spanish-language skills than do Spanish-speaking LM children with weaker initial English skills?).

It was hypothesized that for all skills, children with higher initial skills in either L1 or L2 would benefit more from an intervention designed to promote skill development in the other language than would children with lower initial L1 or L2 skills. Moreover, it was hypothesized that for those skills that are specific-language independent (i.e., PA), results would support the CUP model, whereas for those skills that are language-specific (i.e., print knowledge, receptive and definitional vocabulary), results would support a CUP-independent DIH and transfer of specific linguistic information.

Method

Participants

Ninety-four Spanish-speaking LM children from 10 classes in a Head Start center in Los Angeles, California, participated in this study. All children participating in this study were born in the United States. Fifty-one (54.3%) participants were boys, and all were Latinos. The mean age of the participants was 54.51 months ($SD = 4.72$ months).

Measures

Measures of emergent literacy skills were administered to children in both English and Spanish using the Preschool Comprehensive Test of Phonological and Print Processing (P-CTOPPP; Lonigan, Wagner, Torgesen, & Rashotte, 2002) and the P-CTOPPP-Spanish (Lonigan, Farver, & Eppe, 2002). The P-CTOPPP contains five subtests: Receptive Vocabulary, Definitional Vocabulary, Blending, Elision, and Print Knowledge. The Vocabulary and Print Knowledge subtests of the P-CTOPPP-Spanish are a

direct translation of the items on the English version of the assessment. The Blending and Elision subtests—both measures of PA—of the P-CTOPPP-Spanish are a Spanish-language adaptation of the Blending and Elision subtests of the P-CTOPPP. The P-CTOPPP was the development version of the Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, & Rashotte, 2007), which includes only versions of the Definitional Vocabulary, Phonological Awareness (a combination of blending and elision items), and Print Knowledge subtests. Subtests on the TOPEL have good evidence of validity, with strong correlations between the TOPEL subtests and other measures of each construct.

Vocabulary measures. On the Definitional Vocabulary subtest, children were shown a picture and then asked to name the object in the picture and to describe one of its important features. This subtest contained 40 items that each had two parts. The first part of this task assessed children's expressive vocabulary skills, whereas the second part assessed children's definitional vocabulary skills. On the Receptive Vocabulary subtest, children were shown a page with four pictures and asked to point to the picture of the thing named by the examiner. This subtest contained 40 items. Internal consistency reliability for Receptive and Definitional Vocabulary subtests in both languages was moderate to high in this sample (Receptive Vocabulary: English $\alpha = .87$, Spanish $\alpha = .83$; Definitional Vocabulary: English $\alpha = .98$, Spanish $\alpha = .97$).

Phonological awareness measures. Items on the Elision subtest required children to remove parts of words to form a new word. Items on the Blending subtest required children to combine words or parts of words to form a new word. The English Blending subtest contained 21 items, and the English Elision subtest contained 18 items. Of the 21 Blending items, nine were multiple choice, and 12 were free response. The 18 Elision items were split evenly between multiple choice and free response. Both the Spanish Blending subtest and the Spanish Elision subtest contained 18 items, with items split evenly between multiple choice and free response. Internal consistency reliabilities for both Blending and Elision subtests in both languages were adequate to marginal in this sample (Blending: English $\alpha = .86$, Spanish $\alpha = .81$; Elision: English $\alpha = .72$, Spanish $\alpha = .66$).

Print knowledge measures. The Print Knowledge subtest assessed children's print concepts, letter discrimination, word discrimination, letter-name knowledge, and knowledge of letter-sound correspondence. The print concepts and letter discrimination items were multiple-choice items in which children were shown a page with four pictures and asked to point to the picture that had letters or that "could be read." Word discrimination items were multiple-choice items in which children were shown a page with four pictures and asked to point to the one that "could be read." Letter-name and letter-sound knowledge items included both multiple-choice and free-response items in which a child was either asked to point to a letter corresponding to the name or sound spoken by the examiner or to name or provide the sound associated with a letter displayed on a page. Both English and Spanish versions of the Print Knowledge subtest contained 36 items (four print concept, letter discrimination, and word discrimination items, 16 letter-name knowledge items, and eight letter-sound correspondence items). Internal consistency reliability was moderate to high for both languages (English $\alpha = .93$, Spanish $\alpha = .88$).

Procedure

Informed consent was obtained from parents of participants prior to participation in the study. Children were administered the P-CTOPPP in both English and Spanish both before and after implementation of the intervention at the beginning (October/November) and end (May/June) of the preschool year. The assessments were administered by bilingual graduate and undergraduate research assistants who were trained in the administration of the P-CTOPPP. These research assistants were not involved in the implementation of the intervention, and they had no knowledge of the intervention conditions to which children were assigned. Administration of the assessments was counterbalanced by language and was done over 2 days for each participant, with each session lasting approximately 20–30 min. Children were spoken to in the language in which the test was being administered and were reminded of which language to use if they responded in the other language. Responses were only coded as correct if they were given in the language being assessed.

Intervention

Children were randomly assigned to one of three intervention conditions. One condition was a business-as-usual control condition in which children received only their classroom curriculum (HighScope Preschool Curriculum; HighScope Educational Research Foundation, 2013). The HighScope Preschool Curriculum takes an approach called *active participatory learning* in which children build knowledge through a learning experience that involves direct interactions with people and objects (see www.highscope.org). Milestones children achieve through this curriculum are aligned with state standards, and teachers use a consistent daily routine and planned environment to deliver instruction. Children are provided the opportunity to make plans on their own and later reflect upon what happened.

The other two conditions involved small-group pull-out instruction in oral language, phonological awareness, and print knowledge that used the activities of the Literacy Express Preschool Curriculum (Lonigan, Clancy-Menchetti, Phillips, McDowell, & Farver, 2005; Lonigan, Farver, Phillips, & Clancy-Menchetti, 2011); children in these conditions also received their classroom curriculum. To improve children's oral language skills, interventionists used small-group instruction with dialogic reading. The dialogic reading techniques involved shared book reading between adults and children in which adults asked children open-ended questions throughout the reading of the book to encourage children to "tell the story" on their own. Questions were initially simple and focused on the pictures in the book. As children's language skills and familiarity with the book improved, questions became more complex, requiring children to describe how pictures and other elements of the book related to each other and to other literary elements, such as plot. To improve children's PA skills, interventionists during small-group activities focused on word games using pictures to help children better understand that words are made up of individual units of sound. Instruction initially focused on large units of sound and progressed to smaller units of sound over the course of the preschool year. To improve children's print knowledge skills, interventionists used activities that were primarily centered on improving children's knowledge of the alphabet. These activities initially involved recognition of the letters in the

children’s names and gradually moved to introduction of the names of all letters as well as the sounds corresponding to the letters.

The intervention lasted 21 weeks. Children participated in the small-group sessions four times each week. Each daily session lasted approximately 20 min. All small-group intervention activities were conducted by four bilingual graduate research assistants. Children in one of the intervention conditions received the small-group pull-out instruction in English only (English-only condition). Children in the other intervention condition (transitional condition) received instruction in Spanish for the first 9 weeks of the intervention. At that point, instruction was transitioned to English. Over a period of 3–4 weeks, each lesson previously given in Spanish was reviewed and delivered in English. After that point, children in the transitional condition received instruction only in English.

Fidelity of intervention implementation. Throughout the intervention, those conducting the intervention kept session attendance logs for children in both intervention conditions, and classroom attendance records were obtained for children in the control condition. Children in both the English-only and the transitional conditions were present for 86% of all sessions. Children in the control condition had an attendance rate of 87%. Each week, interventionists’ small-group sessions were observed and rated by the intervention supervisor using a 5-point scale of fidelity of implementation (e.g., activities conducted in prescribed manner, content of session, pacing; 1 = *low fidelity*, 5 = *high fidelity*). Across interventionists, 90%–98% of the rated sessions received a score of 5, indicating that the intervention was provided to children as intended.

Results

Descriptive statistics for both the intervention and control conditions at pretest and posttest are shown in Tables 1 and 2, respectively. To provide a basis for comparison of this sample to other samples, we converted responses on the P–CTOPPP to TOPEL scores. Children’s scores were in the low-average to

below-average range on the Definitional Vocabulary ($M = 77.88$, $SD = 17.01$), PA ($M = 80.96$, $SD = 11.44$), and Print Knowledge ($M = 90.46$, $SD = 10.32$) subtest equivalents of the TOPEL at pretest. Zero-order correlations within skill, both within and across languages and time points, are shown in Table 3. English and Spanish Print Knowledge skills were correlated within language, across time points, and across languages. Similarly, English and Spanish PA measures were correlated within language as well as across languages both within and across time points. Receptive and Definitional Vocabulary skills were significantly correlated within languages. English Receptive Vocabulary skills at pretest and at posttest were significantly correlated with Spanish Receptive Vocabulary skills at posttest but not Spanish Receptive Vocabulary skills at pretest.

Regression analyses were used to examine whether skills in one language at pretest moderated the impact of the intervention on the measure of the same construct in the other language at posttest in two separate intervention condition contrasts (i.e., English-only intervention condition vs. control condition; transitional intervention condition vs. control condition). Because vocabulary knowledge, PA, and print knowledge are distinct skills, analyses were conducted for each outcome separately. We examined both L1 to L2 transfer and L2 to L1 transfer. In these analyses, multiple regression models were conducted with three steps. For the analyses examining L1 to L2 transfer, the first step included the main effect of intervention condition as well as both L1 and L2 pretest skills. In the second step, an Intervention Condition \times Initial L1 Skill interaction term was added to the models. In the third step, an Intervention Condition \times Initial L2 Skill interaction term was added to the models. For the analyses examining L2 to L1 transfer, the first step included the main effect of intervention condition as well as both L1 and L2 pretest skills. In the second step, an Intervention Condition \times Initial L2 Skill interaction term was added to the models. In the third step, an Intervention Condition \times Initial L1 Skill interaction term was added to the models. We probed significant interactions by evaluating the simple effects of intervention condition at one standard deviation above and one

Table 1
Descriptive Statistics for Control and Intervention Conditions on Emergent Literacy Skills in Both English and Spanish at Pretest

Outcome	Min–max possible	Intervention condition		
		Control Adjusted <i>M</i> (<i>SD</i>)	English only Adjusted <i>M</i> (<i>SD</i>)	Transitional Adjusted <i>M</i> (<i>SD</i>)
Child age (months)		54.41 (5.56)	54.00 (4.19)	55.26 (3.78)
English measures				
Receptive Vocabulary	0–40	22.63 (6.26)	23.41 (7.33)	24.32 (5.45)
Definitional Vocabulary	0–80	26.78 (17.28)	30.08 (18.00)	35.69 (13.22)
Blending	0–21	9.62 (3.36)	9.71 (4.34)	10.10 (4.22)
Elision	0–18	4.23 (1.91)	5.29 (2.72)	5.36 (2.89)
Print Knowledge	0–36	10.29 (6.84)	11.52 (6.99)	13.68 (6.02)
Spanish measures				
Receptive Vocabulary	0–40	21.80 (5.34)	20.26 (4.47)	19.53 (6.67)
Definitional Vocabulary	0–80	17.91 (14.61)	22.87 (17.34)	17.76 (15.99)
Blending	0–18	8.26 (3.09)	8.22 (2.98)	8.40 (4.19)
Elision	0–18	3.66 (1.73)	4.27 (2.14)	3.38 (1.77)
Print Knowledge	0–36	7.99 (5.50)	9.99 (5.80)	10.55 (7.86)

Note. $N = 94$. Adjusted M = scores adjusted for chronological age. Min = minimum; max = maximum.

Table 2
Descriptive Statistics for Control and Intervention Conditions on Emergent Literacy Skills in Both English and Spanish at Posttest

Outcome	Min–max possible	Intervention condition		
		Control Adjusted <i>M</i> (<i>SD</i>)	English only Adjusted <i>M</i> (<i>SD</i>)	Transitional Adjusted <i>M</i> (<i>SD</i>)
English measures				
Receptive Vocabulary	0–40	28.33 (5.63)	30.62 (5.85)	31.79 (3.95)
Definitional Vocabulary	0–80	41.23 (16.85)	47.45 (12.96)	52.28 (12.07)
Blending	0–21	12.69 (3.51)	14.31 (3.33)	14.43 (3.04)
Elision	0–18	6.37 (1.51)	7.96 (3.24)	8.04 (3.51)
Print Knowledge	0–36	16.61 (7.96)	20.11 (9.01)	23.90 (7.56)
Spanish measures				
Receptive Vocabulary	0–40	23.79 (4.03)	24.58 (4.07)	27.03 (5.74)
Definitional Vocabulary	0–80	25.74 (15.97)	25.90 (19.30)	32.66 (18.30)
Blending	0–18	10.59 (3.02)	11.13 (2.49)	12.71 (4.06)
Elision	0–18	5.52 (1.32)	5.94 (1.75)	7.40 (2.95)
Print Knowledge	0–36	12.83 (6.28)	13.14 (6.62)	16.54 (8.90)

Note. *N* = 94. Adjusted *M* = scores adjusted for chronological age. Min = minimum; max = maximum.

standard deviation below the mean of the moderator (Cohen & Cohen, 1983). All continuous variables included in regressions were mean centered prior to analyses.

English-Only Intervention Condition Contrasts

English-language outcomes. Results for the analyses that contrasted the English-only intervention condition and the control condition are shown in Table 4. All initial skills measured in English were significant unique predictors of English-language posttest scores, whereas only initial-Spanish scores for Blending and Print Knowledge measures were significant unique predictors of posttest skills measured in English. Consistent with the previously reported impact analysis (Farver et al., 2009), there was a significant main effect of intervention condition such that children exposed to the intervention scored higher than did children in the control condition on all English-language outcomes. In the second step of the regressions, there were significant moderation effects for both initial-Spanish-Receptive-Vocabulary and initial-Spanish-Elision scores. In the third step of the regressions, none of the initial-English-skill interaction terms were significant; however, the addition of the initial-English-Elision interaction term caused the initial-Spanish-Elision interaction term to become nonsignificant, suggesting that children transferred a CUP about PA that was not language specific. In contrast, when the initial-English-Receptive-Vocabulary interaction term was added to the model,

the initial-Spanish-Receptive-Vocabulary interaction term remained a significant unique predictor, suggesting that children transferred specific linguistic information across languages, supporting a CUP-independent DIH.

Results of analyses probing the significant interactions are shown in Figure 1. For both the Receptive Vocabulary and Elision outcomes, the simple effects of the intervention were significant at high levels of initial Spanish skill but not at low levels of initial skill (high Receptive Vocabulary: $\beta = .41, p < .01$; low Receptive Vocabulary: $\beta = .01, p = .97$; high Elision: $\beta = .50, p < .001$; low Elision: $\beta = .10, p = .38$).

Spanish-language outcomes. As shown in Table 4, there were significant main effects of all initial Spanish skills on Spanish-language outcomes. There were no significant main effects of initial English skills or intervention condition on Spanish-language outcomes. No Intervention Condition \times Initial-English Skill interaction term significantly predicted Spanish-language outcomes in Step 2 of the regression, and no Intervention Condition \times Initial-Spanish Skill interaction term significantly predicted Spanish-language outcomes in Step 3 of the regressions.

Transitional Intervention Condition Contrasts

English-language outcomes. Results of the analyses that contrasted the transitional intervention and the control conditions are shown in Table 5. All initial skills measured in English

Table 3
Correlations Among Emergent Literacy Skills Between Languages Within and Across Observations

Measure	English Pretest– English Posttest	Spanish Pretest– Spanish Posttest	English Pretest– Spanish Pretest	English Posttest– Spanish Posttest	English Pretest– Spanish Posttest	Spanish Pretest– English Posttest
Receptive Vocabulary	.67**	.71**	.12	.23*	.21*	.08
Definitional Vocabulary	.81**	.91**	–.02	.16	.06	.08
Elision	.72**	.54**	.35**	.52**	.43**	.31**
Blending	.64**	.70**	.39**	.46**	.24*	.49**
Print Knowledge	.81**	.83**	.70**	.73**	.65**	.67**

Note. *N* = 94.

* $p < .05$. ** $p < .01$.

Table 4
Standardized Regression Coefficients From Final Models and Change in R² for Pretest Main Effects and Interaction Terms for the English-Only Versus Control Condition Comparisons

Predictor variables	Outcome									
	Receptive Vocabulary		Definitional Vocabulary		Elision		Blending		Print Knowledge	
	ΔR ²	B	ΔR ²	B	ΔR ²	B	ΔR ²	B	ΔR ²	B
English-language outcomes										
Step 1	.49***		.65***		.59***		.52***		.75***	
Intervention condition		.20*		.19*		.30**		.23*		.17*
English pretest		.67***		.77***		.53***		.61***		.62***
Spanish pretest		.08		.11		.11		.21*		.28**
Step 2	.04*		.00		.04*		.01		.00	
Condition × Spanish Pretest		.21*		−.01		.16		−.07		−.06
Step 3	.00		.00		.01		.01		.00	
Condition × English Pretest		−.05		−.04		.08		−.11		.04
Spanish-language outcomes										
Step 1	.61***		.91***		.42***		.47***		.68***	
Intervention condition		.09		.00		.17		.10		.03
Spanish pretest		.77***		.95***		.52***		.66***		.77***
English pretest		.08		.00		.12		.01		.08
Step 2	.00		.00		.02		.00		.00	
Condition × English Pretest		.06		.05		.20		.07		.14
Step 3	.00		.00		.02		.02		.02	
Condition × Spanish Pretest		−.02		.02		−.16		−.16		−.17

Note. N = 63.
* p < .05. ** p < .01. *** p < .001.

were significant unique predictors of their respective English skills at posttest. Initial Spanish Blending and Spanish Print Knowledge scores were significant unique predictors of English Blending and English Print Knowledge at posttest, respectively. Additionally, all main effects of intervention condition significantly predicted English-language outcomes such that children in the transitional intervention condition had higher posttest scores than did children in the control condition for all English-language outcomes. None of the interaction terms involving initial-Spanish skills added in the second step of the regression significantly predicted children’s English-language outcomes. When the Intervention Condition × Initial-English-Skill interaction terms were added in the third step of the regression, children’s initial English Blending skills moderated the effect of the intervention for English Blending outcomes.

Results of the analysis probing the significant Blending interaction are shown in Figure 2. At high levels of initial English Blending skills, the simple effect of intervention condition was not significant ($\beta = .06, p = .61$). At low levels of initial English Blending skills, the simple effect of intervention condition was significant ($\beta = .47, p < .001$).

Spanish-language outcomes. As shown in Table 5, all initial skills measured in Spanish were significant unique predictors of their respective Spanish skills at posttest. In addition, all main effects of intervention condition significantly predicted children’s Spanish-language outcomes such that children in the transitional intervention condition had higher Spanish-language skills at posttest than did children in the control condition. When the Intervention Condition × Initial-English Skill interaction terms were added to the models in the second step of the regression, there were

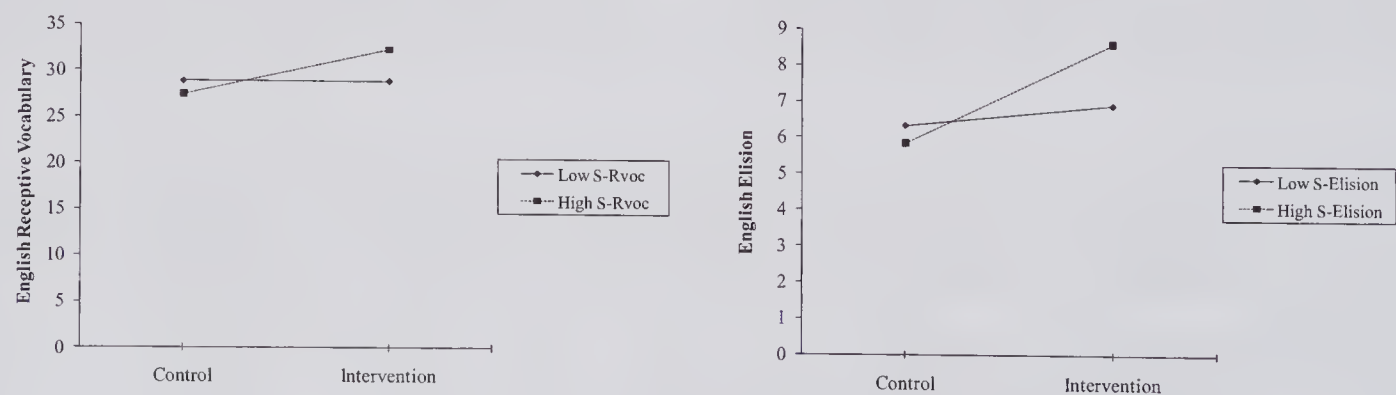


Figure 1. Adjusted posttest scores for English Receptive Vocabulary (Figure 1a) and English Elision (Figure 1b) outcomes for children with varying levels of initial skill on Spanish Receptive Vocabulary (S-Rvoc; Figure 1a) and Spanish Elision (S-Elision; Figure 1b) measures for English-only intervention condition comparison.

Table 5

Standardized Regression Coefficients From Final Models and Change in R^2 for Pretest Main Effects and Interaction Terms for the Transitional Versus Control Condition Comparisons

Predictor variable	Outcome									
	Receptive Vocabulary		Definitional Vocabulary		Elision		Blending		Print Knowledge	
	ΔR^2	B	ΔR^2	B	ΔR^2	B	ΔR^2	B	ΔR^2	B
English-language outcomes										
Step 1	.55***		.74***		.59***		.57***		.82***	
Intervention condition		.34***		.31***		.30**		.27**		.38***
English pretest		.61***		.73***		.58***		.52***		.56***
Spanish pretest		-.05		.07		.00		.36***		.25**
Step 2	.01		.00		.01		.00		.01	
Condition \times Spanish Pretest		.12		-.03		.05		.02		-.13
Step 3	.00		.00		.02		.03*		.00	
Condition \times English Pretest		-.07		.06		.14		-.20*		.04
Spanish-language outcomes										
Step 1	.63***		.82***		.52***		.62***		.78***	
Intervention condition		.32***		.18**		.30**		.29**		.20**
Spanish pretest		.77***		.86***		.59***		.76***		.80***
English pretest		.08		.08		.15		-.09		.07
Step 2	.00		.02*		.07**		.00		.00	
Condition \times English Pretest		.06		.13*		.19*		-.04		.11
Step 3	.00		.00		.07**		.00		.00	
Condition \times Spanish Pretest		-.01		-.03		.27**		.02		-.10

Note. $N = 63$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

significant moderating effects of both initial-English Definitional Vocabulary and initial-English Elision skills. When the Intervention Condition \times Initial-Spanish-Skill interaction terms were added to the models in the third step of the regression, there was a significant moderating effect of initial-Spanish Elision skills on Spanish Elision outcomes. Both of the significant interactions from the second step remained significant when tested in the context of the interaction terms added in the third step, suggesting that children transferred specific linguistic information across languages for Definitional Vocabulary, supporting a CUP-independent DIH, and both specific and common linguistic information across languages for Elision, supporting both a CUP-independent DIH and a CUP model.

Results of analyses probing the significant interactions are shown in Figure 3. For Definitional Vocabulary, the simple effect of intervention condition was significant at high initial levels of English Definitional Vocabulary ($\beta = .31, p < .001$) but not at low initial levels of Definitional Vocabulary ($\beta = .04, p = .62$). Similarly, the simple effect of intervention condition was significant at high initial levels of English Elision ($\beta = .60, p < .001$) but not at low initial levels of English Elision ($\beta = .01, p = .92$). Results probing the significant interaction of Spanish Elision skills are shown in Figure 4. At high initial levels of Spanish Elision skills, the simple effect of intervention condition was significant ($\beta = .57, p < .001$). At low initial levels of Spanish Elision skills, the simple effect of intervention condition was not significant ($\beta = .02, p = .87$).

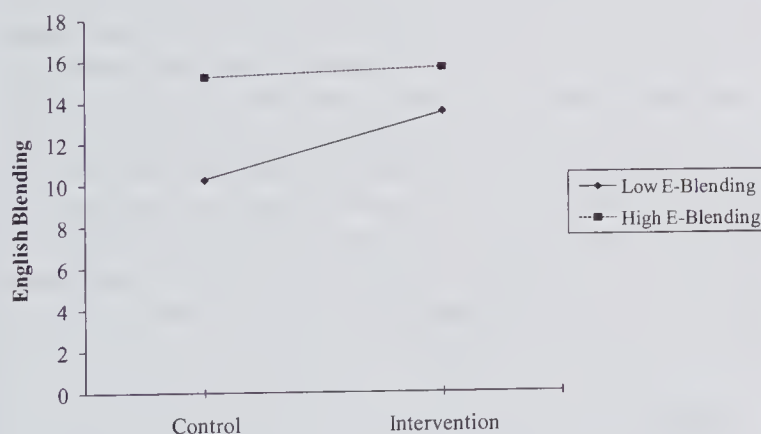


Figure 2. Adjusted posttest scores for English Blending outcomes for children with varying levels of initial skill on English Blending measure (E-Blending) for transitional intervention condition comparison.

Discussion

In this study, we evaluated the presence and type of transfer of emergent literacy skills from one language to another for Spanish-speaking LM preschool children. Beyond findings of co-occurrence of skills in L1 and L2, which are the data typically taken to demonstrate that transfer has occurred, we evaluated whether children's skills in one language would facilitate gains in the other language when children were exposed to an effective intervention. That is, this study addressed the question of whether providing the context in which transfer could occur (i.e., an effective intervention) allowed it to occur. Overall, the results of the study suggest a limited role of transfer in the development of emergent literacy skills for Spanish-speaking preschool LM children. We reasoned that if transfer from L1 to L2 (or vice versa) occurred, children with more skills in one language would show

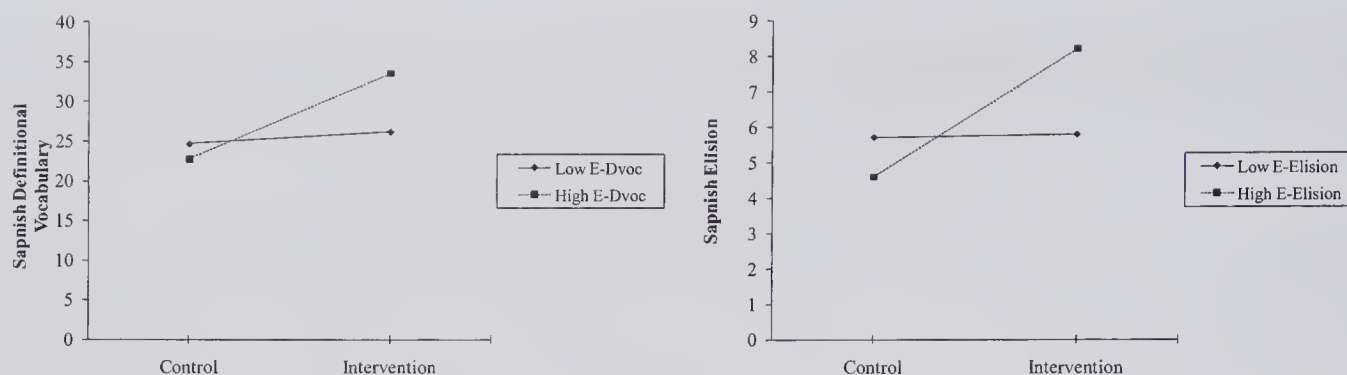


Figure 3. Adjusted posttest scores for Spanish Definitional Vocabulary (Figure 3a) and Elision (Figure 3b) outcomes for children with varying levels of initial skill on English Definitional Vocabulary (E-DV; Figure 3a) and Elision (E-Elision; Figure 3b) measures for transitional intervention condition comparison.

greater gains in the other language as a result of the intervention than children with fewer skills in that language because they would have more skills that they were capable of transferring across languages. Furthermore, this study addressed whether these effects represented transfer of specific linguistic information (a CUP-independent DIH) across languages or represented language-independent transfer (i.e., CUP). We reasoned that for language-independent skills such as PA, a CUP across languages would account for cross-linguistic relations and that for language-dependent skills such as vocabulary and print knowledge, transfer of specific linguistic information across languages would occur. Results provided partial support for these predictions for vocabulary and PA.

In contrast to most prior studies that have attempted to study transfer of skills in samples of LM children, which have used correlational analyses (e.g., Leafstedt & Gerber, 2005; López & Greenfield, 2004; Tabors et al., 2003), this study evaluated transfer in the context of an experimental study of an effective intervention (Farver et al., 2009). A significant positive correlation between a skill in L1 and L2 does not provide strong evidence of transfer because the source of the positive correlation could be due to multiple possible factors, of which transfer is just one. For instance, a positive correlation could be the result of common strong or weak learning environments for the skill in both L1 and L2. Alternatively, a positive correlation could reflect the degree to

which children's general cognitive abilities facilitate acquisition of the skill in both L1 and L2. By experimentally manipulating instruction in this study, we were able to examine the degree to which ability level in L1 influenced learning in L2. Additionally, although most prior research has considered only the possibility of L1 to L2 transfer, there is no reason to expect that emergent literacy skills cannot also transfer from L2 to L1 (Dressler & Kamil, 2006). Therefore, this study examined these relations as well.

In this study, children's print knowledge and PA were correlated across languages, but children's vocabulary skills were generally not correlated across languages. It was expected that PA skills would be significantly correlated across languages because PA is language-independent to the extent that sounds are the same across languages. It was also expected that print knowledge would be significantly correlated across languages because although print knowledge is a language-specific skill, it is relatively similar across English and Spanish as visual representations of many letters are identical and the sounds that correspond to these letters are often the same across languages (although names for the letters differ across languages). The finding that vocabulary skills were not consistently correlated across languages was not surprising. Aside from cognates, vocabulary knowledge is language-specific. In fact, studies indicate that vocabulary knowledge for LM children is often not significantly correlated across languages (e.g., Gottardo & Mueller, 2009).

Results of this study did not support a broad role for transfer in the acquisition of emergent literacy skills. The findings demonstrated that 15 of 20 possible effects of the intervention (i.e., comparing treatment conditions to the control condition on outcomes in both languages) were significant. The only intervention effects that were not statistically significant were the effects for Spanish-language outcomes when comparing the English-only intervention condition—where there was no instruction in Spanish—with the control condition. Of the 15 significant intervention effects, only four cross-language Initial Skill \times Intervention Condition interactions were significant.

Vocabulary

Results for vocabulary outcomes indicated that Spanish-speaking LM children transferred specific linguistic information about vocabulary across languages. Children with higher initial

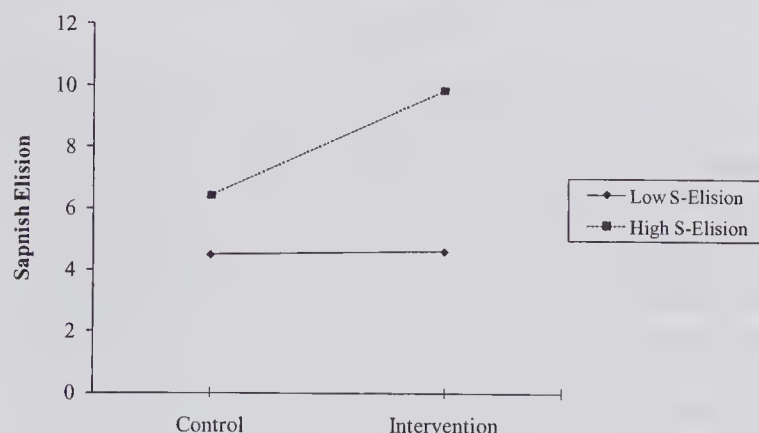


Figure 4. Adjusted posttest scores for Spanish Elision outcomes for children with varying levels of initial skill on Spanish Elision measure (S-Elision) for transitional intervention condition comparison.

vocabulary knowledge in one language benefitted more from the intervention on vocabulary outcomes in the other language than did children with lower initial vocabulary knowledge. These children were exposed to adequate amounts of instruction in English and Spanish to allow their prior Spanish and English vocabulary knowledge to facilitate the acquisition of new vocabulary knowledge in the language of instruction. For both significant vocabulary interactions, the inclusion of a second interaction term with pretest skills in the same language as the outcome did not diminish the unique predictive value of the initial interaction term, indicating that the moderating effect was specific to children's vocabulary in the language that was different from the outcome. Because vocabulary knowledge is not a general skill but is specific knowledge of words in a child's lexicon, it appears that children are able to capitalize on this knowledge of and familiarity with words they have in one language to learn words in another language.

We hypothesized that any measure that significantly moderated the impact of the intervention for one intervention condition contrast (e.g., English-only intervention condition vs. control condition) would do so in the other intervention condition contrast (e.g., transitional intervention condition vs. control condition); however, this was not the pattern of results obtained. There are several possible explanations for why the findings for vocabulary measures were inconsistent with one another. Prior research has suggested that L1 and L2 oral language skills are entirely separate constructs (Gottardo & Mueller, 2009) and that L1 and L2 oral language skills often are not correlated with one another or are even negatively correlated (e.g., Tabors et al., 2003). In this study, neither Receptive Vocabulary nor Definitional Vocabulary skills were significantly correlated across languages at pretest; however, transfer effects were found despite the apparent lack of a relation between L1 and L2 vocabulary knowledge at pretest. The varying languages of instruction across intervention conditions may partially account for the inconsistent results between contrasts. Definitional Vocabulary is a more complex measure than Receptive Vocabulary, requiring children to be able both to name objects and to describe a feature of the object. It is possible that children in the English-only intervention condition did not have the background knowledge in Spanish about these objects necessary to demonstrate transfer of this more complex skill. Children in the transitional intervention condition were exposed to instruction designed to improve their oral language skills in both Spanish and English, facilitating the development of knowledge about objects independent of language and allowing children to transfer knowledge from one language to another. Although this may explain why children in the English-only intervention condition did not transfer definitional vocabulary skills across languages, it does not explain why children in the transitional intervention condition did not transfer receptive vocabulary skills across languages. It does not appear that the overall impact of the intervention on these skills can help explain these results, however, as overall effect sizes of the intervention for both receptive and definitional vocabulary skills were of similar magnitude for both conditions (see Tables 4 and 5, and Farver et al., 2009). Furthermore, it does not appear that zero-order correlations in this study between initial vocabulary knowledge and vocabulary knowledge at posttest can provide insight into this finding, as the cross-language, cross-time relations between receptive vocabulary skills were of similar magnitude to the cross-language, cross-time relations of definitional vocabulary skills.

Although these explanations could provide insight as to how children's vocabulary skills transfer from one language to another, additional research is necessary to determine which, if any, of these explanations is most likely to explain the cross-language transfer of children's vocabulary skills.

Phonological Awareness

Children with higher initial elision skills in one language benefitted more from the intervention on elision outcomes in the other language than did children with lower initial elision skills. For comparisons of the English-only versus control intervention conditions, results supported language-independent transfer, whereas for comparisons of the transitional versus control intervention conditions, results supported both the transfer of specific linguistic information and transfer due to a CUP. We hypothesized that results for PA outcomes would support only language-independent transfer (i.e., transfer due to a CUP). Although there is no clear explanation for these inconsistent findings, the varying results could be an artifact of differing languages of instruction across intervention conditions (i.e., children in the transitional condition received instruction in both English and Spanish). Children in the transitional intervention condition benefitted from a CUP and transferred specific linguistic information across languages for Spanish outcomes. It was expected that children would be able to utilize their CUP across languages to transfer knowledge about PA from English to Spanish and vice versa; however, this was not what the results indicated. It is unclear why children who received instruction in both Spanish and English would transfer information about PA to Spanish, but not to English.

These results indicated that the high correlations between PA measured in Spanish and PA measured in English found in some studies (e.g., Dickinson et al., 2004; López & Greenfield, 2004; Tabors et al., 2003) may not represent the result of the transfer of Spanish PA skills to English PA skills. Rather, such correlations may reflect the development of an underlying PA ability that is not language dependent. The results of Branum-Martin et al. (2006), however, indicated that there are components of PA that are unique to each language. That is, they found that PA measured in English and PA measured in Spanish were best represented as distinct factors, despite a high correlation between the two factors.

In their lexical restructuring model (LRM), Metsala and Walley (1998) proposed that as children's vocabularies grow, their mental representation of words undergoes restructuring from a holistic form to a more fine-grained, segmented form of words. The LRM, along with the general abilities model of transfer (Castilla et al., 2009) and the notion of a CUP (Cummins, 1981), can help merge the results of this study with the findings of Branum-Martin et al. (2006) that there are components of PA that are unique to each language. Lexical restructuring occurs at the local level (i.e., only for words that a child knows). As this shift in children's mental representation of words takes place, they have better access to component parts of words. With this increased access comes the possibility of the development of PA skills. Some of this knowledge about the sounds of words is language independent, and children are able to detect and manipulate the word sounds of both of their languages, as evidenced by this study's partial support for language-independent transfer. Additionally, a recent study by Atwill, Blanchard, Christie, Gorin, and García, (2010) suggested

that receptive vocabulary skills among LM children moderated the relation between L1 and L2 PA skills, such that correlations between L1 and L2 PA skills were lower for children with lower L1 receptive vocabulary than for children with higher L1 receptive vocabulary. This finding provides further support for the theory that some knowledge of the sounds of words is language independent; however, some of this knowledge about the sounds of words is specific to those words that resulted in lexical restructuring and does not lead to increased PA skills in another language, as evidenced by the finding that there are components of PA that are unique to each language (Branum-Martin et al., 2006) and this study's partial support for transfer of specific linguistic information.

Because both elision and blending are presumably measures of the same underlying PA construct, it was expected that if cross-language transfer effects were found for elision skills they would also be found for blending skills; however, this was not the case. Elision is a more difficult task than is blending for preschool children (Anthony et al., 2011; Lonigan, Burgess, Anthony, & Barker, 1998). It is possible that children who had higher initial blending skills were near the ceiling of the measure and did not have as much room to show substantial improvement on the measure as did children with higher initial elision skills. Although mean scores at pretest and posttest were substantially higher for the Blending subtests than for the Elision subtests, these scores were not approaching the ceiling of the measure, ruling this out as an explanation for the lack of transfer of blending skills. The effects of the intervention were smaller overall for blending than for elision (see Tables 4 and 5; Farver et al., 2009), suggesting that there might have been a partial ceiling effect on blending.

Print Knowledge

Initial print knowledge skills in either L1 or L2 also did not moderate the impact of the intervention for any intervention-condition contrast. The Print Knowledge subtest of the P-CTOPPP is mostly knowledge of letter names and letter-sound correspondence. Letters in Spanish and English are mostly the same, but they have different names and several make different sounds. This is similar to vocabulary because objects are the same across languages but are described using different words and sounds in each language. There were significant transfer effects for vocabulary knowledge but not for print knowledge. Other studies have claimed that aspects of print knowledge (e.g., letter-sound knowledge) demonstrate cross-language transfer (e.g., Lindsey et al., 2003); however, these studies simply examined cross-language correlations. The high cross-language correlations between print knowledge skills at pretest in this study (see Table 3) may indicate that cross-language transfer of print knowledge skills already occurred for these children prior to the intervention. For example, if the children in this study had already been exposed to activities that increase print knowledge skills in both languages, there would have been only limited information about print knowledge that these children could transfer across languages as a result of the intervention.

Limitations and Future Research

Despite the advantages of an experimental design for examining cross-language transfer, this study contained several limitations

that point to potential directions for future research in the area of cross-language transfer of emergent literacy skills. First, this study had a small sample size and was relatively underpowered to detect moderation effects. Future studies that use larger samples when examining cross-language transfer of emergent literacy skills in the context of an experimental design study may uncover additional evidence of transfer. Second, these analyses did not fully address the issue of cross-language transfer, as transfer may be a phenomenon that occurs over a longer period of time than the duration of this study, and future research could make use of longitudinal designs to examine the cross-language transfer of emergent literacy skills. These results solely suggest that cross-language transfer may occur when children are exposed to activities specifically designed to improve their emergent literacy skills. To address the question of transfer in the absence of targeted instruction would require a longitudinal study. Longitudinal designs could also help determine the point during development at which cross-language transfer is most likely to occur and can help to inform instruction.

Summary and Conclusions

Results of this study supported only a limited role for the transfer of emergent literacy skills for Spanish-speaking preschool LM children. Although results of prior correlational studies have indicated that children's literacy and preliteracy skills in more than one language are interdependent, this study suggested that only certain skills transfer from one language to another. Prior correlational studies do not address whether transfer of these skills from one language to another occurs because they cannot rule out alternative explanations, such as environments that support the development of skills in both languages simultaneously. This study, in which an experimental manipulation of instruction was used to evaluate potential transfer of these skills, provides partial evidence of language-independent transfer (i.e., transfer due to a CUP) and the transfer of specific linguistic information, depending on the outcome evaluated. This study further advances the knowledge of the relations between L1 and L2 emergent literacy skills for LM children by examining cross-language transfer of emergent literacy skills through the experimental manipulation of instruction, which is a novel method of examining this issue.

Support for the transfer of specific linguistic information as evidenced by the moderation effect of initial vocabulary knowledge (for both English-only and transitional intervention condition comparisons) and elision skills (for the transitional intervention condition comparisons) is a sort of nontraditional Matthew effect (Stanovich, 1986), or a *Mateo effect*. Matthew effects imply that children who need instruction the least are able to benefit from it the most (i.e., the rich get richer). However, the Matthew effect presumably would not occur across languages, as children with higher initial skills in one language do not necessarily have higher initial skills in their other language, as evidenced by the typical finding that children's L1 and L2 vocabulary knowledge are not correlated or are negatively correlated with one another. The finding that this effect does occur across languages is unique to this study.

The significant moderation effects of elision skills (for the English-only intervention condition comparison) that partially supported language-independent transfer represent a more traditional

version of the Matthew effect, in which children with greater underlying ability in one language benefit from instruction in that language to a greater extent than do children with less underlying ability. Differences in task demands can account for the varying results seen for vocabulary knowledge and PA skills. PA tasks are skill-based and require children to manipulate the individual sound components of words, whereas vocabulary assessments are not general skill-based tasks; rather, they draw upon knowledge of specific words. For vocabulary knowledge, children learn new words that they may already know in their other language. Children can then capitalize on their conceptual knowledge of vocabulary (i.e., specific-language-independent vocabulary knowledge; Bedore, Peña, García, & Cortez, 2005) and apply it to their L2. For PA skills, children simply build on a foundation of knowledge about what PA is generally, rather than build on knowledge that is specific to one language (as is the case with vocabulary) and apply this concept to increasingly difficult tasks. This pattern of results was obtained for the English-only intervention condition comparisons but not for the transitional intervention condition comparisons, suggesting that language of instruction may play a role in the transfer of specific linguistic information across languages.

References

- Anthony, J. L., Solari, E. J., Williams, J. M., Schoger, K. D., Zhang, Z., Branum-Martin, L., & Francis, D. J. (2009). Development of bilingual phonological awareness in Spanish-speaking English language learners: The roles of vocabulary, letter knowledge, and prior phonological awareness. *Scientific Studies of Reading, 13*, 535–564. doi:10.1080/10888430903034770
- Anthony, J. L., Williams, J. M., Durán, L. K., Gillam, S. L., Liang, L., Aghara, R., . . . Landry, S. H. (2011). Spanish phonological awareness: Dimensionality and sequence of development during the preschool and kindergarten years. *Journal of Educational Psychology, 103*, 857–876. doi:10.1037/a0025024
- Atwill, K., Blanchard, J., Christie, J., Gorin, J. S., & García, H. S. (2010). English-language learners: Implications of limited vocabulary for cross-language transfer of phonemic awareness with kindergarteners. *Journal of Hispanic Higher Education, 9*, 104–129. doi:10.1177/1538192708330431
- Atwill, K., Blanchard, J., Gorin, J. S., & Burstein, K. (2007). Receptive vocabulary and cross-language transfer of phonemic awareness in kindergarten children. *Journal of Educational Research, 100*, 336–346. doi:10.3200/JOER.100.6.336-346
- Bedore, L., Peña, E., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference? *Language, Speech, and Hearing Services in Schools, 36*, 188–200. doi:10.1044/0161-1461(2005/020)
- Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*, 170–181. doi:10.1037/0022-0663.98.1.170
- Castilla, A. P., Restrepo, M. A., & Perez-Leroux, A. T. (2009). Individual differences and language interdependence: A study of sequential bilingual development in Spanish–English preschool children. *International Journal of Bilingual Education and Bilingualism, 12*, 565–580. doi:10.1080/13670050802357795
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*, 222–251.
- Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: A theoretical framework* (pp. 3–49). Sacramento, CA: California State Department of Education.
- Cummins, J. (1991). Interdependence of first- and second-language proficiency in bilingual children. In E. Bialystok (Ed.), *Language processing in bilingual children* (pp. 70–89). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511620652.006
- Cummins, J. (2008). Teaching for transfer: Challenging the two solitudes assumption in bilingual education. In J. Cummins & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 5. Bilingual education* (2nd ed., pp. 65–75). New York, NY: Springer Science + Business Media.
- Cummins, J., Swain, M., Nakajima, K., Handscombe, J., Green, D., & Tran, C. (1984). Linguistic interdependence among Japanese and Vietnamese immigrant students. In C. Rivera (Ed.), *Communicate competence approaches to language proficiency assessment: Research and application* (pp. 60–81). Clevedon, England: Multilingual Matters.
- Dickinson, D. K., McCabe, A., Clark-Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of phonological awareness in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics, 25*, 323–347. doi:10.1017/S0142716404001158
- Dressler, C., & Kamil, M. (2006). First- and second-language literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 197–238). Mahwah, NJ: Erlbaum.
- Farver, J. A. M., Lonigan, C. J., & Eppe, S. (2009). Effective early literacy skill development for young Spanish-speaking English language learners: An experimental study of two methods. *Child Development, 80*, 703–719. doi:10.1111/j.1467-8624.2009.01292.x
- Fitzgerald, J. (1995). English-as-a-second-language learners' cognitive reading processes: A review of research in the United States. *Review of Educational Research, 65*, 145–190.
- Gottardo, A. (2002). The relationship between language and reading skills in bilingual Spanish–English speakers. *Topics in Language Disorders, 22*, 46–70. doi:10.1097/00011363-200211000-00008
- Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology, 101*, 330–344. doi:10.1037/a0014320
- Hemphill, F. C., Vanneman, A., & Rahman, T. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress. Statistical Analysis Report*. Washington, DC: Institute of Education Sciences, National Center for Education Statistics.
- HighScope Educational Research Foundation. (2013). *HighScope preschool curriculum*. Available from <http://www.highscope.org>
- Leafstedt, J. M., & Gerber, M. M. (2005). Crossover of phonological processing skills: A study of Spanish-speaking students in two instructional settings. *Remedial and Special Education, 26*, 226–235. doi:10.1177/07419325050260040501
- Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology, 95*, 482–494. doi:10.1037/0022-0663.95.3.482
- Lonigan, C. J., Burgess, S. R., Anthony, J. L., & Barker, T. A. (1998). Development of phonological sensitivity in two- to five-year old children. *Journal of Educational Psychology, 90*, 294–311. doi:10.1037/0022-0663.90.2.294
- Lonigan, C. J., Clancy-Menchetti, J., Phillips, B. M., McDowell, K., & Farver, J. M. (2005). *Literacy express: A preschool curriculum*. Tallahassee, FL: Literacy Express.

- Lonigan, C. J., Farver, J. M., & Eppe, S. (2002). *Preschool Comprehensive Test of Phonological & Print Processing: Spanish Version (P-CTOPPP-S)*. Tallahassee, FL: Authors.
- Lonigan, C. J., Farver, J. A. M., Phillips, B. M., & Clancy-Menchetti, J. (2011). Promoting the development of preschool children's emergent literacy skills: A randomized evaluation of a literacy-focused curriculum and two professional development models. *Reading and Writing, 24*, 305–337. doi:10.1007/s11145-009-9214-6
- Lonigan, C. J., Schatschneider, C., & Westberg, L. (2008). Impact of code-focused interventions on young children's early literacy skills. In *Developing early literacy: Report of the National Early Literacy Panel* (pp. 107–151). Washington, DC: National Institute for Literacy.
- Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2002). *Preschool Comprehensive Test of Phonological and Print Processing*. Tallahassee, FL: Authors.
- Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *Test of Preschool Early Literacy*. Austin, TX: ProEd.
- López, L. M., & Greenfield, D. B. (2004). The cross-language transfer of phonological skills of Hispanic head start children. *Bilingual Research Journal, 28*, 1–18. doi:10.1080/15235882.2004.10162609
- Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in Grades K–2 in Spanish-speaking English-language learners. *Learning Disabilities Research & Practice, 19*, 214–224. doi:10.1111/j.1540-5826.2004.00107.x
- Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ: Erlbaum.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–406. doi:10.1598/RRQ.21.4.1
- Tabors, P. O., Pérez, M., & Lopez, L. M. (2003). Dual language abilities of bilingual four-year olds: Initial findings from the early childhood study of language and literacy development of Spanish-speaking children. *NABE Journal of Research and Practice, 1*, 70–91.
- U.S. Census Bureau. (2007). *Language use in the United States: 2007*. Retrieved November 29, 2012, from <http://www.census.gov/hhes/socdemo/language/data/acs/appendix.html>
- U.S. Census Bureau. (2011). *The Hispanic population in the United States: 2011*. Retrieved November 29, 2012, from <http://www.census.gov/population/hispanic/data/2011.html>
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology, 30*, 73–87. doi:10.1037/0012-1649.30.1.73
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., . . . Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology, 33*, 468–479. doi:10.1037/0012-1649.33.3.468
- Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development, 69*, 848–872.

Received January 30, 2012

Revision received January 3, 2013

Accepted January 4, 2013 ■

Enhancing a Brief Writing Intervention to Combat Stereotype Threat Among Middle-School Students

Natasha K. Bowen, Kate M. Wegmann, and Kristina C. Webber
University of North Carolina at Chapel Hill

Experimental research has demonstrated both the deleterious effects of negative stereotypes about ability on academic performance and the relative ease with which stereotypes can be countered in educational settings. The extent to which stereotypes contribute to the achievement gap between American students from dominant social and economic groups and students from other groups is not precisely known, but the potential of brief, inexpensive interventions targeting stereotype threat to reduce the gap is worthy of further examination. Although researchers studying brief social psychological interventions sometimes mention the importance of the context in which interventions occur, they have not included manipulations of the environment in their interventions. In the current experimental study, a test of the effects of a brief self-affirming writing assignment was conducted in a new sample of middle-school students ($n = 132$), and an environmental enhancement to the writing exercise was tested ($n = 274$). Consistent with previous findings, the self-affirming intervention reduced the average decline in Social Studies grades over the school year compared with a neutral condition (effect size, $ES, .57$). The combination of the affirming writing assignment with an environmental enhancement had superior effects to the writing assignment alone ($ES .53$).

Keywords: brief intervention, stereotype threat, middle school, social environment, academic performance

Stereotype threat is defined as “the threat of being viewed through the lens of a negative stereotype or the fear of doing something that would inadvertently confirm that stereotype” (Steele, 2003, p. 111) and the resulting negative effects on performance. Social psychological research has established the power of negative stereotypes about ability to impede the academic performance of students from stereotyped groups, such as African Americans, Latinos, students from low-income families, and women (Croizet & Claire, 1998; McKown & Weinstein, 2003; Nguyen & Ryan, 2008; Steele, 1997; Steele & Aronson, 1995). Recent studies indicate that in addition to causing underperformance on academic tasks among stigmatized groups, stereotype threat also impedes the learning process (Mangels, Good, Whiteman, Maniscalco, & Dweck, 2012; Rydell, Shiffrin, Boucher, Van Loo, & Rydell, 2010; Taylor & Walton, 2011)—a finding that suggests the potential for even greater harm for the targets of negative stereotypes.

Research on mechanisms of stereotype threat in specific performance situations has revealed how easily threat can be triggered, for example, by framing activities as tests of ability or reminding students of their own stereotyped demographic characteristics. Much research examining the processes by which stereotype threat

causes underperformance focuses on “acute protective reactions” (Steele, 2003, p. 124), that is, immediate psychological responses through which individuals attempt to maintain self-integrity in the face of threat. Acute reactions to self-integrity posed by stereotypes include the emotional, cognitive, and physiological elements of anxiety (Mangels et al., 2012), efforts to suppress or deny stereotypes (Logel, Iserman, Davies, Quinn, & Spencer, 2009), and efforts to disprove or prevent the fulfillment of stereotypes (Schmader, Johns, & Forbes, 2008; Steele, 2003; Taylor & Walton, 2011). Researchers suggest that these acute reactions cause a diversion of cognitive resources that would otherwise be committed to the “controlled attention, effortful processing, and active self-regulation” (Schmader et al., 2008, p. 342) required for optimal performance in academic situations.

Stereotype Threat Intervention

Fortunately, experimental studies with secondary and postsecondary students demonstrate how countering the psychological processes that interfere with performance can be surprisingly simple and effective. Encouraging a brief focus on self-affirmation before a stressful exam, for example, can have a significant positive effect on performance. Perhaps the simplest and least expensive intervention studied so far involves asking students to write a 15-min essay about a positive value that is important to them (Cohen, Garcia, Apfel, & Master, 2006; Cohen, Garcia, Purdie-Vaughns, & Brzustoski, 2009; Miyake et al., 2010; Taylor & Walton, 2011). Not only did the exercise improve performance on immediate academic tasks in some studies (Taylor & Walton, 2011), in others it improved course grades weeks or even years later (Cohen et al., 2006, 2009; Miyake et al., 2010). Of most

This article was published Online First December 17, 2012.

Natasha K. Bowen, Kate M. Wegmann, and Kristina C. Webber, School of Social Work, University of North Carolina at Chapel Hill.

This study was funded with a Jane H. Pfouts Research Grant Award from the School of Social Work, University of North Carolina at Chapel Hill.

Correspondence concerning this article should be addressed to Natasha K. Bowen, School of Social Work, University of North Carolina at Chapel Hill, 27599–3550. E-mail: nbowen@email.unc.edu

relevance to the current study, African American seventh graders who wrote a brief self-affirming essay early in the year in their Social Studies classrooms obtained better Social Studies grades over the grading term as well as better grade-point averages in general. Results persisted through the end of eighth grade (Cohen et al., 2009).

The possibility that the academic performance of stereotyped students could be improved with simple, brief interventions is as tantalizing as it is hard to believe. Even well-researched, school-based interventions requiring extensive resources and staff effort rarely obtain substantial effect sizes (Powers, 2005; Yeager & Walton, 2011), and many schools are not equipped to buy or implement such interventions even if they were available (Powers, Bowen, & Bowen, 2010). Yeager and Walton (2011), however, present possible explanations for the success of brief interventions in general. They refer, for example, to the concept in social psychology that “every attitude and behavior exists in a complex field of forces” (p. 274), some of which promote and some of which impede the learning or performance of individual students. Brief interventions may work by removing cognitive defenses that serve as “critical barriers” (impeding forces) to performance that prevent some students’ use of opportunities (promotive forces) in their own minds or in the classroom (Yeager & Walton, 2011, p. 275). The potentially large effects of brief interventions that have been observed may require the preexistence of appropriate learning opportunities and other promotive forces, such that once critical, subjective psychological barriers are removed, better performance can occur. The long-lasting effects of brief interventions may also require the existence of cognitive and environmental elements supportive of positive “recursive processes” (Cohen et al., 2009) once barriers are removed. Cohen et al. (2009) invoked this type of process to explain how effects of their seventh-grade intervention persisted through eighth grade, but they appear to refer primarily to recursive cognitive processes in the student. In the classroom, we suspect a central requirement would be the presence of a teacher who is responsive to information about positive characteristics of a student and/or signs of improved performance of a student.

Social-Environmental Nature of Stereotype Threat

Although discussions on brief social psychological interventions sometimes include mention of the context or environment in which the interventions occur, brief interventions so far have only attempted to manipulate individual-level psychological factors. A more ecological approach, such as that articulated by Bronfenbrenner (1979, 2005) and routinely used in the discipline of social work, also involves the social environment as a target of interventions. This perspective is also supported by the phenomenological variant of ecological systems theory (Spencer, 1999), which illustrates how experiences in the environment (e.g., stereotypes, teacher expectations), students’ self-perceptions, and coping strategies interact to affect outcomes. From these perspectives, stereotype threat does not exist solely within the psychology of an individual, but at the intersection of the individual and his or her environment (Shapiro & Neuberg, 2007). Interventions to counter threat therefore could target individual factors, such as cognitive defenses, and/or characteristics in the social environment. In school settings, for example, characteristics of the classroom con-

text for learning would be considered potentially influential intervention targets.

In the microsystem (Bronfenbrenner, 1979) of the classroom, the teacher is an important environmental force. Teachers dictate, model, incentivize, enforce, and reinforce expectations and norms for child and youth behavior in the classroom. The influence of teachers extends to the operation of stereotypes in the classroom (Jussim & Harber, 2005; Martens, Johns, Greenberg, & Schimel, 2006; Steele, 2003). Teachers have the power (although not always the training or support) to construct classroom environments in which all students know they are valued, cared about, and safe from negative stereotypes; in which they experience competence and self-efficacy; and in which they are expected to progress and succeed academically regardless of background characteristics (Markus, Steele, & Steele, 2000). In the absence of complete or accurate and relevant personal information, teachers and other school staff may resort to stereotypes to form judgments of students (Guyll, Madon, Prieto, & Scherr, 2010). In the case of African American, Latino, and Native American students, the stereotypes by which they might be judged could include being low achievers, not being as intelligent as other students, and not having adequate language skills. Unfortunately, such stereotypes may contribute to lowered teacher expectations for students (Gonzalez & Ayala-Alcantar, 2008; Thomas, Caldwell, Faison, & Jackson, 2009), which are particularly powerful predictors of later achievement for negatively stereotyped students (Hinnant, O’Brien, & Ghazarian, 2009; McKown & Weinstein, 2008). Because teachers’ attitudes and beliefs about students significantly affect the classroom environment experienced by each student (Eccles & Roeser, 2011; Goodenow, 1993), including the nature of peer interactions (Farmer, Lines, & Hamm, 2011), and students’ subsequent performance (Raudenbush, 1984), teachers’ attitudes and beliefs are a potentially key environmental characteristic to target in stereotype reduction interventions.

The current study had two goals: (a) to replicate the findings of earlier studies (Cohen et al., 2006, 2009) that demonstrated positive effects of a brief, self-affirming writing intervention on the grades of African American middle-school students, and (b) to examine the effects of an enhanced intervention that simultaneously targeted the classroom environment. On the basis of the original study by Cohen et al. (2006), the first goal involved comparing the effects on grades of a self-affirmation writing condition and a neutral writing condition. We hypothesized, based on the 2006 study, that writing a self-affirming essay would positively affect middle-school students’ grades. The second goal involved comparing the effects of writing a self-affirming essay with the effects of writing a self-affirming essay that was also read by a teacher. The enhancement was simple, feasible, and consistent with the focus in both social psychology and social work on the interaction between individual psychology and the social environment. We hypothesized that if a student’s teacher read his or her self-affirming essay, the positive student-level psychological effects of the writing exercise would be enhanced. Specifically, teachers who became aware of their students’ positive values and experiences might develop less stereotyped views of students, more positive expectations of students, and higher regard for students—in other words, they might see students more as individuals and less as stereotypes. This change in the social environment in turn would be expected to reinforce or amplify the

achievement-boosting cognitive “recursive processes” initiated by the writing exercise alone (Cohen et al., 2006, 2009).

Method

Sample

The current analysis focuses on a subset of students who took part in a larger study comparing the grades of middle-school students assigned to six brief writing conditions. All regular education students ($N = 585$) in Grades 6, 7, and 8 at one middle school in a mostly urban area of the southeastern United States took part in the study. The intervention took place while students were in homeroom with their homeroom teachers ($n = 24$). The homeroom teacher of each student also taught one of the student’s core subjects. Students in the larger study were African American ($n = 407$), Latino ($n = 117$), or “other” ($n = 61$; White, Asian, multiracial) according to school records. (Academic performance of students in the multiracial category was higher than that of African American and Latino students, so they were grouped with the Asian and White sample members for analyses.)

In the current analysis, we focus on African American and “other” race/ethnic students in four conditions ($n = 313$). Latino sample members were excluded in order to focus on the primary race/ethnic groups in the original studies (Cohen et al., 2006, 2009). Of the 313 African American and “other” race/ethnic students, 274 had one or more observations on their quarterly Social Studies grade data and could be included in the analysis. Thirty-nine students with no Social Studies grade data were excluded. These students did not differ from the 274 with grade data with respect to gender ($\chi^2 = .037, p = .847$), race/ethnicity ($\chi^2 = .065, p = .799$), grade level ($\chi^2 = 3.76, p = .153$), or condition ($\chi^2 = 5.279, p = .152$).

Conditions did not differ significantly by gender ($\chi^2 = 6.65, p = .084$), race/ethnicity ($\chi^2 = 1.78, p = .619$), or grade level ($\chi^2 = 3.815, p = .702$). Prior year standardized math and reading test scores were available at the individual level for about 60% of the sample. Mean student test performance in math and reading before the intervention also did not differ significantly across conditions ($F = 1.33, p = .265$; $F = .208, p = .891$, respectively). These tests strongly suggest that the random assignment process resulted in equivalent groups across which the intervention effects can be evaluated, a point that is especially important because our first outcome measurement (Quarter 1 grades) occurred after the intervention.

Table 1 presents the demographic characteristics of the students included in the current analyses. Table 1 also indicates that 80% of the students at the school as a whole participated in the federal school lunch program, and a majority were performing below grade level in math, reading, or both (North Carolina Department of Public Instruction, 2011). These characteristics of the sample suggest that virtually all students at the school belonged to at least one negatively stereotyped group and/or had their own history of low performance. Therefore, unlike the Cohen et al. (2006, 2009) studies, our study does not include an adequately sized dominant cultural comparison group (e.g., White, middle class). Our interest is in promoting academic excellence of all students rather than making the performance of one dominant group the standard by which others are judged (Hilliard, 2003). Also, the current study

Table 1

Characteristics of Analysis Sample and the School Population Overall

Individual-level variable	Percent (n)	School-level characteristics ^a
African American	86.1 (236)	80% school lunch program participation <40% at or above grade level in reading <50% at or above grade level in math (individual-level data not available on these variables)
Asian/White/multiracial	13.9 (38)	
Boys	49.6 (136)	
Girls	50.4 (138)	
Grade 6	34.7 (95)	
Grade 7	34.3 (94)	
Grade 8	31.0 (85)	
Total N	274	

^a Source: North Carolina Department of Public Instruction (2011).

targeted sixth-, seventh-, and eighth-grade students, rather than just seventh graders as in Cohen et al.’s studies.

Procedure

Students were randomly assigned to write either a self-affirming essay or a neutral essay. The 24 homeroom teachers of the students were randomly assigned to either read or not read the essays of their students. See Table 2 for more detail on the conditions. Students completed the exercise in their homeroom classrooms at the beginning of the school day 2 weeks before the end of the first grading period. Teachers were provided with instructions and a script to follow when introducing and distributing the envelopes. Teachers and students were unaware of the nature of the experiment, its hypotheses, or students’ assignment to conditions. Teachers were aware that students responded to different prompts, but did not know the purpose of the prompts. Each student received an envelope labeled only with his or her name. Each envelope contained self-explanatory instructions, the assigned prompt, and paper for the essay. Students were not provided any information regarding whether or not their essays would be read by the teacher; however, students were explicitly told that their essays would not be graded. Teachers were advised not to look at students’ writing during the exercise or to talk about the writing exercise after it was completed. After 15 min of writing, students replaced all materials in their envelopes, sealed them, and returned them to teachers. Depending on whether they were assigned to the reading condition or not, teachers either returned their students’ essays unread to the study coordinator at the school or read them. The \$25 teacher incentive for taking part in the study was doubled for teachers assigned to the reading condition.

Measures

Students were given codes according to their randomly assigned intervention condition. At the end of the school year, gender, race/ethnicity, grade level, and quarterly grades were linked to student codes by school staff and given (without names) to the researchers. Quarterly grades were recorded on a 100-point scale. Consistent with previous studies of the effects of the writing intervention in middle school (Cohen et al., 2006, 2009), the current study reports on intervention effects on Social Studies grades. The coding of gender and race/ethnicity reflected each

Table 2
Description of Study Prompts and Conditions

Prompt	Teacher does not read essay (<i>n</i> = 12 teachers)	Teacher reads essay (<i>n</i> = 12 teachers)	<i>n</i>
Neutral: Think about the following list of values. Choose the one that is least important to you. Write for 15 minutes about why this value that is not very important to you might be important to someone else.	NEUTRAL Control condition (<i>n</i> = 74)	NEUTRAL + TEACHER Environmental only (<i>n</i> = 67)	141
Self-affirming: Think about a value, belief, talent, or skill you have that you are proud of. Write for 15 minutes about this positive part of yourself and why it is important to you.	AFFIRMATION Psychological only (<i>n</i> = 58)	AFFIRMATION + TEACHER Psychological + Environmental Intervention (<i>n</i> = 75)	133
Cases included in analysis (with outcome data)	132	142	274

variable’s role in the each hypothesis test according to guidelines provided by Singer and Willet (2003, p. 115). In tests of both hypotheses, gender was a control variable. Because it was not of substantive interest, it was coded as a centered dummy. In tests of the first hypothesis (the self-affirming essay would positively affect students’ Social Studies grades), race/ethnicity was dummy coded with “other” as the reference category. This coding facilitated comparisons of the effects of conditions on the performance of African American and “other” students (i.e., White, multiracial or Asian), consistent with the original study (Cohen et al., 2006). Like gender, race/ethnicity was centered as a control variable for tests of the second hypothesis (teachers’ reading of essays would enhance effects of the self-affirming essays). Conditions were represented in these tests by three dummy variables, with the self-affirming essay alone (no teacher reading of essays) as the reference category. This coding allowed us to precisely test our hypotheses in relation to the original studies, but did not permit comparisons of other pairs of conditions. Values for the variable time (Academic Quarter 1, 2, 3, 4) were recoded to (time – 1), so the intercept of equations was zero.

Analyses

Analyses were conducted with Stata/SE version 10.0 (StataCorp LP, 1985–2007). For each hypothesis, we tested a series of longitudinal hierarchical linear models (HLM) using maximum likelihood estimation to examine how writing conditions affected trajectories of middle-school students’ quarterly Social Studies grades. First an unconditional means model was estimated, followed by an unconditional growth model. Time (quarter) was modeled at Level 1; individual students were modeled at Level 2. Quadratic effects of time were also examined by adding a squared term. Condition and demographics were then entered. Because students had their homeroom teacher for one core subject, but other teachers in the sample for other subjects, and because the condition variable accounted partially for information in the teacher variable, no third-level clustering was modeled. Main effects of condition variables in the models represented condition effects on the intercept of students’ grade trajectories. Effects of Time × Condition product terms represented condition effects on the slope of trajectories. Random effects of statistically significant Level 1 variables (time and interaction terms including time) were tested using Stata’s likelihood ratio test and retained if significant. Because the covariance between the random effects of slope and

intercept was nonsignificant, the default diagonal error matrix was modeled instead of an unstructured matrix. All two-way interactions between pairs of predictors (time, gender, race/ethnicity, and condition) were examined and retained only if statistically significant.

In the notation used by Singer and Willett (2003), the following equations were estimated:

$$Y_{ij} = \pi_{0i} + \varepsilon_{ij}$$
 (Unconditional means equation, Level 1). (1)

$$Y_{ij} = \pi_{0i} + \pi_{1i}(\text{Time}) + \pi_{2i}(\text{Time})^2 + \varepsilon_{ij}$$
 (Unconditional growth equation, Level 1). (2)

$$\pi_{0i} = \gamma_{00} + \gamma_{01}(\text{Condition}) + \gamma_{02}(\text{gender}) + \gamma_{03}(\text{race/ethnicity}) + \gamma_{04}(\text{Level 2 interactions}) + \zeta_{0i}$$
 (Conditional Level 2 equation predicting the intercept of the Level 1 equation). (3)

$$\pi_{1i} = \gamma_{10} + \gamma_{11}(\text{cross-level interactions}) + \zeta_{1i}$$
 (Conditional Level 2 equation predicting the slope of the Level 1 equation). (4)

In Equations 1 and 2, the dependent variable, Y_{ij} , is the Social Studies grade of an individual student (*i*) for a quarter (*j*). π_{0i} is the mean Social Studies for an individual (*i*) across the four time points, and ε_{ij} is the difference of the individual’s score at any time point from the mean of his or her score across all time points. The term π_{1i} in Equation 2 is the mean effect of time (school year quarter) on the Social Studies score of an individual (*i*). A second unconditional growth model tested for quadratic effects of time (π_{2i}). The squared term was removed when it was not significant. Similarly, the random effects of time were tested at this step and removed if not significant. In Equation 3, the dependent variable is the intercept of the unconditional models (π_{0i}). It is predicted by a mean intercept across individuals (γ_{00}), Level 2 predictors, and a term for the deviation of each individual’s score from his or her predicted score. γ_{01} to γ_{03} are regression coefficients for condition, gender, and race/ethnicity. The term γ_{04} represents coefficients for a sequence of three 2-way interactions between the Level 2 predictors, which were tested one at a time and retained or omitted depending on their significance. In Equation 4, the dependent variable is the slope term of the Level 1 growth equation

(π_{1i}). It is predicted by a mean slope across individuals (γ_{10}), a series of two-way cross-level interactions created by multiplying time by each of the Level 2 predictors, and a deviation score for each individual.

Effect sizes of statistically significant effects were calculated using the Hedge's g formula for HLM models with cluster-level assignment presented by What Works Clearinghouse (2008, Appendix B):

$$g = \frac{\gamma}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}$$

If the intervention had a main effect on the outcome, the numerator was the regression coefficient for the main effect. When cross-level interactions were involved, the numerator was the simple slope of the effect. Variance values in the denominator were random variance values of the constant in unconditional intercept models if the intervention had a main effect on the outcome; they were random variance values of time in Level 1 (time only) models if the intervention affected change over time, or the slope, of the outcome trajectory.

Results

Research Question 1: Can the beneficial effects of the self-affirming writing intervention that have been observed in prior research (Cohen et al., 2006, 2009) be replicated with a new sample? The first sequence of analyses compared the Social Studies grade trajectories of students who wrote self-affirming essays (AFFIRMATION) with the grade trajectories of those in the control condition, who wrote neutral essays (NEUTRAL). Neither of the "teacher reads essay" conditions was included in this analysis. The overall mean of Social Studies grades of the 132 students in the two conditions examined was 85.58. The mean unconditional quarterly decline in grades for students in the two conditions was .79 points. Table 3 presents the estimates from the final model.

The table indicates that although condition did not significantly affect the starting level of the grade trajectories for students in the two groups, writing a self-affirming essay did reduce the decline in grades over time relative to the NEUTRAL condition from 1.26 points to .19 points per quarter. The Hedge's g effect size for the

slope difference is .57 (simple slope = 1.07; n for students in NEUTRAL = 74, n in AFFIRMATION = 58; variances of slope in unconditional model for NEUTRAL and AFFIRMATION conditions = 3.79 and 3.18, respectively). Figure 1 illustrates the effect. Effects of the intervention did not differ by race/ethnicity, and at no time point did the levels of grades differ significantly across condition.

Research Question 2: Does the reading of students' self-affirming essays by teachers enhance the positive effects of the writing intervention? The second sequence of analyses compared the Social Studies grade trajectories of students who wrote self-affirming essays (AFFIRMATION) with those of students in the control condition (NEUTRAL) and the two conditions in which teachers read student essays (AFFIRMATION + TEACHER, NEUTRAL + TEACHER). We were particularly interested in the comparison of the AFFIRMATION + TEACHER and AFFIRMATION conditions. The overall mean of Social Studies grades of the 274 students in the four conditions was 86.95. The mean decline in grades in the unconditional model was .55 points per quarter. Table 4 presents estimates from the final model.

The table reveals that the AFFIRMATION + TEACHER condition had a statistically significant impact on the starting level of Social Studies grade trajectories relative to the AFFIRMATION intervention. Two weeks after the writing intervention, students who wrote self-affirming essays and had their essays read by teachers received first-quarter grades that were almost 3.7 points higher than those who wrote affirming essays that were not read by teachers. The Hedge's g effect size for this value was 0.53 (main effect = 3.66; n for AFFIRMATION = 58, n for AFFIRMATION + TEACHER = 75; variances of intercepts in unconditional model for AFFIRMATION and AFFIRMATION + TEACHER conditions = 62.62 and 37.33, respectively). The AFFIRMATION + TEACHER condition did not have a statistically significant effect on the *change* in Social Studies grades over time, relative to the AFFIRMATION condition, only on the starting point. The slope effect of the self-affirming essay (AFFIRMATION) relative to the neutral essay (NEUTRAL) that was related to Hypothesis 1 was the only other significant finding in the test of Hypothesis 2. That is, students in the AFFIRMATION condition did not have significantly higher grades relative to the NEUTRAL or NEUTRAL + TEACHER conditions, and their grades did not decline at a slower

Table 3
Final Model Comparing Effects of Writing a Self-Affirming Essay With Writing a Neutral Essay (With No Teacher Reading of Essays)

Predictor	Unconditional means model	Unconditional growth model	Final model
Intercept	85.58 ($p = .000$)	86.76 ($p = .000$)	87.89 ($p = .000$)
Time		-0.79 ($p = .002$)	-0.19 ($p = .613$)
NEUTRAL effect on intercept (vs. AFFIRMATION)			2.47 ($p = .064$)
African American (vs. White, multiracial, "other")			-2.56 ($p = .176$)
NEUTRAL effect on slope (vs. AFFIRMATION)			-1.07 ($p = .033$)
Centered control			
Gender			-3.30 ($p = .009$)
Random effects			
Intercept	51.33	44.27	40.56
Slope		3.57	3.42
Residual	30.59	23.72	23.57

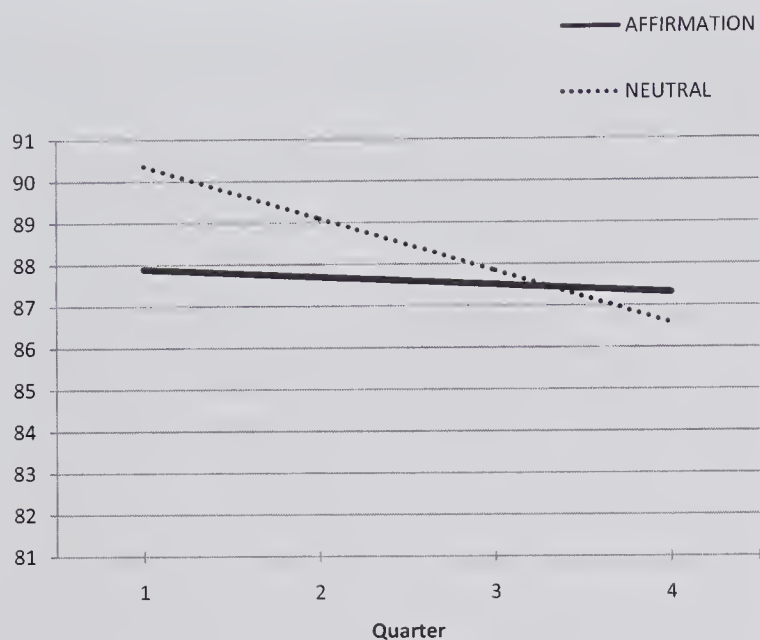


Figure 1. Statistically significant slope effect of condition on quarterly Social Studies grades controlling for gender and race/ethnicity.

rate than those of students in the NEUTRAL + TEACHER condition. The Hedge’s *g* of this effect was presented above. Figure 2 illustrates the significant findings.

Discussion

In this experimental study, we first tested the hypothesis that writing a self-affirming essay would have beneficial effects relative to writing a neutral essay on the Social Studies grades of stereotyped middle-school students. As expected from the previous experimental studies (Cohen et al., 2006, 2009), our sample benefited from the self-affirming essay. The effect was on the change in grades over the school year. Specifically, the typical decline in Social Studies grades that occurred over the school year among students at the target school was significantly slowed among students who wrote affirming essays relative to students randomly assigned to write neutral essays. Because the original Cohen et al.

(2006) study did not test for differences in the rates of change in grades over time, it is not clear whether the higher scores observed in that study among stereotyped students who wrote self-affirming essays was due to abrupt statistically significant increases in grades or a different rate of change in grades over time that led eventually to significantly higher scores. The duration of our study (one academic year) was not long enough to evaluate whether the observed trajectory would extend far enough into the next school year to ultimately lead to statistically higher grades among students who wrote self-affirming essays.

Unlike in the Cohen et al. (2006, 2009) studies, we did not find significant differences in the effects for African American and non-African American students (European American in the Cohen et al. studies). At least two likely explanations exist for this finding. First, our comparison sample of Asian, European American, and multiracial students was small relative to the African American group, giving us limited statistical power to detect differences between African Americans and these groups. Second, because of high rates of poverty and low performance among students at the school, it is likely that many in the non-African American group were subject to other stereotypes, such as those based on social class. Therefore, they did not, on average, represent a comparison group unaffected by stereotypes.

Our findings expand on previous work by using more sophisticated statistical modeling techniques, including sixth and eighth graders in addition to seventh graders, and by having a predominantly low-income sample. The findings suggest that the benefits of writing self-affirming essays in the classroom generalize beyond the characteristics of the original sample to low-income students and students who may belong to more than one stereotyped group. In addition, although it appears that in the first year of the original study (as described in Cohen et al., 2009) students experienced additional booster interventions, our study suggests that one “dose” in the fall can have benefits that last until the end of the school year. Our findings are promising in that they suggest even historically low-performing schools may have promotive forces available to support student success once critical barriers are addressed. They also suggest that positive recursive processes may be “jump-started” in such schools with brief interventions.

Table 4
Final Model Comparing Effects of Writing a Self-Affirming Essay With Writing a Neutral Essay and Two Conditions in Which Teachers Read Essays

Predictor	Unconditional means model	Unconditional growth model	Final model
Intercept	86.95 (<i>p</i> = .000)	87.79 (<i>p</i> = .000)	85.74 (<i>p</i> = .000)
Time		−0.55 (<i>p</i> = .001)	−0.19 (<i>p</i> = .587)
NEUTRAL effect on intercept (vs. AFFIRMATION)			2.38 (<i>p</i> = .057)
NEUTRAL + TEACHER effect on intercept (vs. AFFIRMATION)			1.87 (<i>p</i> = .146)
AFFIRMATION + TEACHER effect on intercept (vs. AFFIRMATION)			3.66 (<i>p</i> = .003)
NEUTRAL effect on slope (vs. AFFIRMATION)			−1.07 (<i>p</i> = .022)
NEUTRAL + TEACHER effect on slope (vs. AFFIRMATION)			−0.05 (<i>p</i> = .923)
AFFIRMATION + TEACHER effect on slope (vs. AFFIRMATION)			−0.23 (<i>p</i> = .614)
Centered controls			
Gender			−3.10 (<i>p</i> = .000)
African American (vs. Asian, White, multiracial)			−3.91 (<i>p</i> = .001)
Random effects			
Intercept	49.76	42.25	36.65
Slope		3.36	3.17
Residual	25.47	19.56	19.55

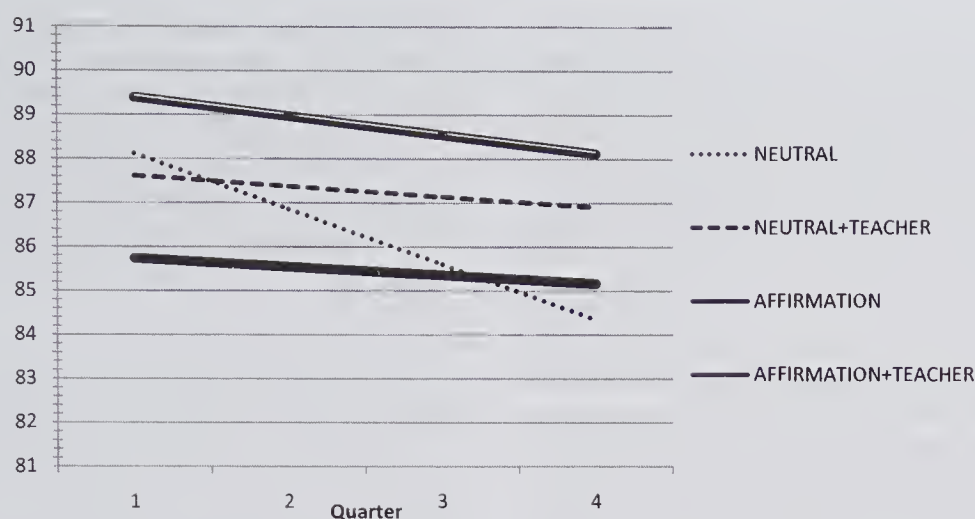


Figure 2. Intercept and slope effects of condition on Social Studies grades controlling for race/ethnicity and gender. The intercept of the AFFIRMATION + TEACHER trajectory is significantly higher than the intercept of the AFFIRMATION trajectory. The slope of the AFFIRMATION trajectory is significantly less steep (downward) than the slope of the NEUTRAL trajectory.

On the basis of principles of social work intervention and an ecological perspective of child development and performance, we expected that by increasing teachers' familiarity with positive values held by students, the beneficial effects of writing a self-affirming essay would be amplified. The recursive psychological processes posited by previous researchers (Cohen et al., 2006, 2009; Yeager & Walton, 2011) would be fueled by any number of teacher gestures and communications toward students, making the classroom environment one in which threats to students' identities would become increasingly less salient. Our second analysis tested this hypothesis. Although we did not directly measure teacher attitudes or behaviors, by randomly assigning students to four conditions, we were able to compare the effects of writing an affirming essay with three other scenarios: writing a neutral essay with and without having it read by a teacher, and writing an affirming essay that is read by a teacher.

Students in the condition receiving the AFFIRMATION + TEACHER intervention had statistically higher Social Studies grades than those in the AFFIRMATION condition (but not those in the NEUTRAL or NEUTRAL + TEACHER conditions). Students in the AFFIRMATION condition, in turn, as in the test of Hypothesis 1, evidenced a slower decline in Social Studies grades over the school year than those in the NEUTRAL condition (but not those in the NEUTRAL + TEACHER condition). The findings indicate that although students can benefit from writing self-affirming essays alone, teachers' reading essays in and of itself does not promote higher achievement. No benefits were observed for students who wrote neutral essays that were read by teachers. It was only when teachers read *self-affirming* essays that our sample of stereotyped students, on average, benefited.

The authors of the original studies of the effects of the self-affirming essay on middle-school students' academic performance (Cohen et al., 2006, 2009) attribute the intervention's beneficial effects to its fueling of positive recursive psychological processes. The act of students writing about positive values they hold affirms positive aspects of their identities, allowing the students to remember their own unique strengths and to diminish the salience of negative stereotypes. Although the existing literature on this inter-

vention attributes its success to self-affirmation and individuation, the writing task itself could also be considered a thought replacement strategy because it requires students to concentrate on something positive rather than any negative thoughts they might be experiencing. Although self-affirmation and individuation are obviously additional cognitive tasks, they do not drain resources from working memory and automatic processing capacity as do other, less beneficial responses to stereotype threat, such as attempting to suppress distressing thoughts or vigilant self-monitoring (Logel et al., 2009; Schmader et al., 2008). Whether cognitive resources are freed or preexisting promotive factors such as motivation are tapped, performance improves, and feedback on the improved performance may lead to future better performance via a snowballing recursive psychological process (Cohen et al., 2009). Finding effects on grades that persisted across the school year even without booster exercises is especially encouraging.

In the current study, assigning teachers to read students' essays has extended individuation to the environmental level through the hypothesis that the essays would help teachers view students as unique individuals rather than relying on preconceived expectations or stereotypes. Knowing unique, personal information about students would allow teachers to provide environmental reinforcement for students' positive identities. The environmental benefit found in teacher reading appears to have amplified the positive effects of the original self-affirmation and individuation strategies.

Our study contributes to the stereotype threat literature by testing an intervention in a population consisting primarily of stigmatized students (e.g., low-income and minority students). Additional strengths of the study include the use of hierarchical longitudinal modeling of grade trajectories, the systematic testing of interactions, the use of controls for gender and race/ethnicity effects on grades, and the calculation of effect sizes. The study used two random assignment procedures—one at the individual level and one at the teacher level. All available evidence suggests that random assignment led to equivalent intervention groups. In the interest of parsimony and because the starting point of our study was a previous study, we focused our hypotheses on comparisons of the previously tested intervention (the self-affirming

essay) with the previously tested control condition and our two new conditions (neutral essay read by teacher and self-affirming essay read by teacher). Therefore, we did not test for differences between every possible pair of conditions. We also focused on Social Studies grades, as was done in the original study. Because Social Studies is considered to be less affected by gender stereotypes (Cohen et al., 2006), intervention effects are presumed to be unconfounded by gender.

A limitation of the study is that we were not able to examine the role of student expectations that their essays would be read by teachers or not. Students were told that their essays would not be graded, but they were not explicitly told whether or not their essays would be read by teachers. We assume that in the absence of statements to the contrary, students expect teachers to read any work they do in the classroom. It is possible, therefore, that simply believing their teachers would read their self-affirming essays enhanced the effects of writing the essays for some students, even if reading the essays caused no change on the part of the teacher. However, if such a cognitive effect occurred for some students, it should have affected those who wrote affirming essays similarly regardless of whether they were in the AFFIRMATION condition or the enhanced AFFIRMATION + TEACHER condition. The finding of different levels of trajectory starting points is inconsistent with this explanation.

An additional limitation is that we did not measure changes in teacher attitudes and behaviors associated with reading student essays. Although we noted a positive effect of teachers' reading affirming essays, future research should examine the exact mechanism by which change occurred. The lack of data on student performance from the prior year is also a limitation, although with random assignment to conditions, this is less of a concern. Our results suggest that the writing intervention on average can benefit students who are low performing, because most of the students in the school were low performing, but we could not test for moderation by prior performance levels.

Implications for Future Research

Future research should seek greater understanding of the experiences of various cultural and economic groups with stereotypes and stereotype threat. The immense variation in cultural, historical, and social environmental experiences that exist within schools and communities across the United States may affect the operation of brief interventions that target psychological and social environmental factors in ways that are currently not fully understood. Effects of this intervention specific to Latino students will be discussed in a forthcoming article.

The interactive perspective suggested in this study is consistent with a typology recently proposed by Shapiro and Neuberg (2007), in which stereotype threat situations are categorized by the roles that both self and others occupy in the threatening situation. For example, students in one cultural group may be more likely to worry about confirming stereotypes to themselves, whereas students belonging to another cultural group may find the threat of confirming stereotypes to other observers to be more salient. Future studies should test hypotheses about the validity and utility of this typology. Better understanding of the nature of threats and the mechanisms by which they are most effectively countered in demographic subgroups will promote the development of appro-

priate brief interventions. If stereotype threats are qualitatively different, coping and compensatory strategies to reduce stereotype threat are likely not "one size fits all"—rather, strategies must be matched appropriately to situational characteristics in order to most effectively negate the threat (Shapiro, 2011). On the basis of their placement in Shapiro's (2011) typology, some students might benefit from writing essays that are not read by teachers or from responding to prompts tailored more to the nature of their experienced threat.

Implications for Practice and Policy

Although it may be tempting to rely on brief, low-cost interventions to improve the academic achievement of stereotyped students, there is a larger lesson to be learned from the current study. Transforming American classrooms into settings free from socially constructed psychological barriers to achievement may be an attainable goal. It is not as daunting a goal as eliminating stereotypes in the greater culture. Teachers can work routinely to create classrooms in which student performance is optimized regardless of the persistence of negative stereotypes in the larger society. Strategies for teachers that are consistent with the brief intervention literature include explicitly countering and debunking stereotypes in the classroom; promoting the growth mindset (Blackwell, Trzesniewski, & Dweck, 2007; Dweck, 2006); building supportive relationships with students; and otherwise consistently affirming the self-integrity, competence, and belonging of each student in the classroom. Having teachers read self-affirming essays is one way to reduce critical psychological barriers to achievement, but combining brief interventions with more sustained and integrative efforts to banish stereotype threat in the classrooms will likely yield the greatest benefits in the long run.

References

- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*, 246–263. doi:10.1111/j.1467-8624.2007.00995.x
- Bronfenbrenner, U. (1979). *The ecology of human development*. Cambridge, MA: Harvard University Press.
- Bronfenbrenner, U. (2005). *Making human beings human: Bioecological perspectives on human development*. Thousand Oaks, CA: Sage.
- Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313*, 1307–1310. doi:10.1126/science.1128317
- Cohen, G. L., Garcia, J., Purdie-Vaughns, V., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science, 324*, 400–403. doi:10.1126/science.1170769
- Croizet, J. C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*, 588–594. doi:10.1177/0146167298246003
- Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.
- Eccles, J. S., & Roeser, R. W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence, 21*, 225–241. doi:10.1111/j.1532-7795.2010.00725.x
- Farmer, T. W., Lines, M. M., & Hamm, J. V. (2011). Revealing the invisible hand: The role of teachers in children's peer experiences.

- Journal of Applied Developmental Psychology*, 32, 247–256. doi:10.1016/j.appdev.2011.04.006
- Gonzalez, R., & Ayala-Alcantar, C. (2008). Critical caring: Dispelling Latino stereotypes among preservice teachers. *Journal of Latinos and Education*, 72, 129–143. doi:10.1080/15348430701828699
- Goodenow, C. (1993). Classroom belonging among early adolescent students: Relationship to motivation and achievement. *The Journal of Early Adolescence*, 13, 21–43. doi:10.1177/0272431693013001002
- Guyl, M., Madon, S., Prieto, L., & Scherr, K. C. (2010). The potential roles of self-fulfilling prophecies, stigma consciousness, and stereotype threat in linking Latino/a ethnicity and educational outcomes. *Journal of Social Issues*, 66, 113–130. doi:10.1111/j.1540-4560.2009.01636.x
- Hilliard, A. G., III. (2003). No mystery: Closing the achievement gap between Africans and excellence. In T. Perry, C. Steele, & A. G. Hilliard, III (Eds.), *Young, gifted, and Black* (pp. 131–165). Boston, MA: Beacon Press.
- Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology*, 101, 662–670. doi:10.1037/a0014306
- Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9, 131–155. doi:10.1207/s15327957pspr0902_3
- Logel, C., Iserman, E. C., Davies, P. G., Quinn, D., & Spencer, S. J. (2009). The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology*, 45, 299–312. doi:10.1016/j.jesp.2008.07.016
- Mangels, J. A., Good, C., Whiteman, R. C., Maniscalco, B., & Dweck, C. S. (2012). Emotion blocks the path to learning under stereotype threat. *Social Cognitive and Affective Neuroscience*, 7, 230–241. doi:10.1093/scan/nsq100
- Markus, H. R., Steele, C. M., & Steele, D. M. (2000). Colorblindness as a barrier to inclusion: Assimilation and nonimmigrant minorities. *Daedalus*, 129, 233–259.
- Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, 42, 236–243. doi:10.1016/j.jesp.2005.04.010
- McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development*, 74, 498–515. doi:10.1111/1467-8624.7402012
- McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology*, 46, 235–261. doi:10.1016/j.jsp.2007.05.001
- Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330, 1234–1237. doi:10.1126/science.1195996
- Nguyen, H-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93, 1314–1334. doi:10.1037/a0012702
- North Carolina Department of Public Instruction. (2011). *School report cards*. Retrieved from <http://www.ncschoolreportcard.org>
- Powers, J. D. (2005). *Evidence-based practice in schools: Current status, potential barriers, and critical next steps* (Doctoral dissertation). University of North Carolina at Chapel Hill.
- Powers, J. D., Bowen, N. K., & Bowen, G. L. (2010). Evidence-based programs in school settings: Barriers and recent advances. *Journal of Evidence Based Social Work*, 7, 313–331. doi:10.1080/15433710903256807
- Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, 76, 85–97. doi:10.1037/0022-0663.76.1.85
- Rydell, R. J., Shiffrin, R. M., Boucher, K. L., Van Loo, K., & Rydell, M. T. (2010). Stereotype threat prevents perceptual learning. *Proceedings of the National Academy of Sciences*, 107, 14042–14047. doi:10.1073/pnas.1002815107
- Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, 115, 336–356. doi:10.1037/0033-295X.115.2.336
- Shapiro, J. R. (2011). Types of threats: From stereotype threat to stereotype threats. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 71–88). New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780199732449.003.0005
- Shapiro, J. R., & Neuberg, S. L. (2007). From stereotype threat to stereotype threats: Implications of a multi-threat framework for causes, moderators, mediators, consequences, and interventions. *Personality and Social Psychology Review*, 11, 107–130. doi:10.1177/1088868306294790
- Singer, J. D., & Willett, J. B., (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
- Spencer, M. B. (1999). Social and cultural influences on school adjustment: The application of an identity-focused cultural ecological perspective. *Educational Psychologist*, 34, 43–57. doi:10.1207/s15326985ep3401_4
- StataCorp LP. (1985–2007). *Stata/SE (Version 10.0)*. College Station, TX: Author.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identify and performance. *American Psychologist*, 52, 613–629. doi:10.1037/0003-066X.52.6.613
- Steele, C. (2003). Stereotype threat and African-American student achievement In T. Perry, C. Steele, & A. Hilliard, III (Eds.), *Young, gifted, and Black* (pp. 109–130). Boston, MA: Beacon Press.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811. doi:10.1037/0022-3514.69.5.797
- Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin*, 37, 1055–1067. doi:10.1177/0146167211406506
- Thomas, O. N., Caldwell, C. H., Faison, N., & Jackson, J. S. (2009). Promoting academic achievement: The role of racial identity in buffering perceptions of teacher discrimination on academic achievement among African American and Caribbean Black adolescents. *Journal of Educational Psychology*, 101, 420–431. doi:10.1037/a0014578
- What Works Clearinghouse. (2008). *WWC procedures and standards handbook, Appendix B—Effect size calculations*. Retrieved from <http://ies.ed.gov/ncee/wwc/help/iddocviewer/Doc.aspx?docId=19&tocId=8#go14>
- Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, 81, 267–301. doi:10.3102/0034654311405999

Received July 10, 2011

Revision received October 31, 2012

Accepted November 13, 2012 ■

A Contextualized View on Long-Term Predictors of Academic Performance

Janine Gut, Giselle Reimann, and Alexander Grob
University of Basel

Several studies show that parents' and teachers' perceptions of children's academic competence are important predictors of children's subsequent academic performance. However, there is a lack of evidence on what precedes these perceptions as well as the dynamics within a comprehensive model. The aim of this study was to investigate the simultaneous influences of child (general intelligence, problem behavior) and family (socioeconomic status, migration background) characteristics as well as parents' and teachers' perceptions of children's competence on children's academic performance in a 3-year longitudinal study with 221 children (52% girls) who were 5 to 7 years of age when they were first measured ($M = 6.15$ years, $SD = 0.80$ years). Structural equation modeling analyses revealed that parents' and teachers' perceptions of children's competence were positively associated with children's general intelligence and negatively associated with family adversity and child problem behavior. Further, parents' perceptions fully mediated the association between family adversity as well as child problem behavior and academic performance. Teachers' perceptions partially mediated the association between child problem behavior and academic performance.

Keywords: family adversity, problem behavior, general intelligence, perceptions of competence, academic performance

Several studies have demonstrated that parents' perceptions of children's competence are important predictors of children's subsequent academic performance (e.g., Phillipson & Phillipson, 2007; Pomerantz & Dong, 2006). Students whose parents hold higher competence perceptions of their offspring receive higher grades, achieve higher scores on standardized tests, and stay longer in school than do those students whose parents hold lower perceptions (Peet, Powell, & O'Donnel, 1997; Phillipson, 2010; Pomerantz & Dong, 2006; Yamamoto & Holloway, 2010). In essence, parents' perceptions of children's competence may act as self-fulfilling prophecies in the way parents communicate perceptions, either implicitly or explicitly, to their children in everyday interactions or provide them with support and opportunities to learn, which in turn may directly or indirectly determine the way children perceive education and perform at school (Chan, 2009; Dandy & Nettelbeck, 2002; Phillipson, 2009).

Although research on parents' perceptions of children's academic competence has yielded important insights into academic outcomes, there is a lack of evidence on what precedes these perceptions, and the aim of the present study was to provide such evidence. Likewise, studies support the inclusion of teachers since teachers' perceptions of children's competence may influence chil-

dren's academic achievement through the provision of positive attention, learning opportunities, and reinforcement of performance (Jussim & Harber, 2005; McLoyd, 1998; Mistry, White, Benner, & Huynh, 2009). In addition, the inclusion of teachers as a valid source of information would reflect behavior and perceptions more closely tied to the educational context and would probably result in a more accurate model because children's home and school settings would be linked (Bronfenbrenner & Morris, 2006; Rutchick, Smyth, Lopoo, & Dusek, 2009). Recognizing these considerations, we designed the present study to investigate associations between child characteristics (general intelligence and problem behavior) and family characteristics (socioeconomic status, family migration background) as predictors of parents' and teachers' perceptions of children's competence and children's subsequent academic performance in a longitudinal framework.

Family Characteristics as Predictors of Academic Performance

Academic achievement and education are powerful markers of career-oriented and financial success as well as of subjective well-being (Friedman, 2005; Mistry et al., 2009). However, ethnic and class-related disparities in educational achievement show that children from families with a migration background and lower socioeconomic status (SES) obtain lower performance scores on mathematics, reading, and science than do children from families without a migration background and with higher SES (e.g., Prenzel et al., 2007). Low-SES families often miss financial resources, which factually is important in the pursuit of higher education. Further, low-SES rates are higher among children of immigrants than among children of nonimmigrants and are highest for young children of immigrants. Immigrant children more often have par-

This article was published Online First February 4, 2013.

Janine Gut, Giselle Reimann, and Alexander Grob, Department of Developmental and Personality Psychology, University of Basel, Basel, Switzerland.

Correspondence concerning this article should be addressed to Janine Gut, Department of Developmental and Personality Psychology, University of Basel, Missionsstrasse 60/62, 4055 Basel, Switzerland. E-mail: janine.gut@unibas.ch

ents with lower educations and limited native language skills (Capps, Fix, Ost, Reardon, & Passel, 2004). Children of immigrants also score lower on measures of parent-child interaction and experience lower levels of cognitive and language stimulation at home at young ages than do nonimmigrant children (Ayoub et al., 2009). Therefore, it is important to consider mediating processes by which family-related risks such as lower SES and a migration background may affect children's subsequent academic performance.

The Mediating Role of Parents' and Teachers' Perceptions of Children's Competence

There are only a few studies that have examined the associations between family demographic variables (e.g., ethnicity, migration status, and SES) and parents' and teachers' perceptions of children's competence. For example, it has been demonstrated that lower SES was associated with lower perceptions of children's competence in parents and teachers (Alexander, Entwisle, & Thompson, 1987; McLoyd, 1998). Research indicates that even in kindergarten and first grade, teachers tend to perceive low-SES children less positively, for example, as having less maturity and fewer self-regulatory skills, than high-SES children (McLoyd, 1998). Such perceptions in teachers may lead them to provide children from families with lower SES with less positive attention, fewer learning opportunities, and less reinforcement. Also, parents under economic pressure may hold lower perceptions of their children's competence on the basis of their own lower educational attainment and lower academic as well as parental self-efficacy beliefs (Elder, Eccles, Ardel, & Lord, 1995).

Child Characteristics as Predictors of Academic Performance

A large number of studies indicate that general intelligence as well as behavioral regulation are significant predictors of children's academic achievement (e.g., Kytälä & Lehto, 2008; McLeod & Kaiser, 2004; Sektnan, McClelland, Acock, & Morrison, 2010; Taub, Keith, Floyd, & McGrew, 2008). In particular, several studies confirmed that general intelligence was one of the best predictors of educational achievement, showing moderate to strong correlations even with variables such as prior school performance, sex, and SES controlled (e.g., Deary, Strand, Smith, & Fernandes, 2007; Harackiewicz, Barron, Tauer, & Elliot, 2002; Kuncel, Hezlett, & Ones, 2004).

On the other hand, children with behavioral difficulties during the first few years of school have been shown to exhibit more problems with academic adjustment and to be at increased risk of dropping out prior to the completion of formal schooling (e.g., Alexander, Entwisle, & Kabbani, 2001; Gutman, Sameroff, & Cole, 2003). However, according to ecological perspectives on child development, academic achievement can be seen as a product of dynamic interactions between individual and environmental factors (Bronfenbrenner, 1979; Bronfenbrenner & Morris, 2006). Therefore, attention needs to be paid to the role of parents and teachers in the association between children's characteristics and academic achievement.

The Mediating Role of Parents' and Teachers' Perceptions of Children's Competence

Ecological perspectives on child development suggest that interactions in the most proximate systems (e.g., home, school) play a significant role in mediating the relationship between child characteristics (e.g., general intelligence, problem behavior) and academic achievement. These mediating processes imply an intervention in the direct associations between children's general intelligence, problem behavior, and academic achievement by home- and school-related factors (e.g., Bronfenbrenner & Morris, 2006). However, only a few studies have investigated whether parents' and teachers' perceptions of children's academic competence may function as mediating processes in the relation between children's general intelligence or problem behavior and children's subsequent academic achievement (Phillipson, 2010; Rutchick et al., 2009).

Regarding children's general intelligence, direct associations with academic performance are well established (e.g., Kuncel et al., 2004). However, studies have also demonstrated that conventional IQ measures typically explain about 25% of the variance in academic achievement and that the predictive value may be increased by a combination with additional variables such as cognitive stimulation at home, achievement goals, or academic self-concept (e.g., Ayoub et al., 2009; Harackiewicz et al., 2002; Kornilova, Kornilov, & Chumakova, 2009). Accordingly, Phillipson (2010) showed that emotional and social support such as expressions of confidence in children's competencies were crucial predictors of children's academic performance in low- as well as high-ability students. However, supportive behavior was lower among parents of low-ability students. Therefore, lower perceptions of children's competencies in parents and teachers of children with lower general intelligence scores may further restrain these children's motivation and achievement in academic-related performance.

Finally, according to the fundamental attribution error (Jones & Harris, 1967), people tend to attribute observed behaviors in other persons to underlying traits rather than to situational influences. In this way, parents and teachers who observe a child engaging in problem behavior may infer that the behaviors reflect the child's underlying dispositions, and they therefore may generally lower their perceptions of the child's abilities to succeed in an academic environment even if the child shows adequate academic performance (McLeod & Kaiser, 2004; McLoyd, 1998; Rutchick et al., 2009). Therefore, children's externalizing behavior problems may have a direct biasing effect on parents' and teachers' perceptions of children's competence as well as an indirect long-term effect on children's subsequent academic achievement.

The Present Study

The aim of the present study was to investigate the simultaneous influences of child (general intelligence, problem behavior) and family (SES, migration status) characteristics as well as parents' and teachers' perceptions of children's competence on children's academic performance in a 3-year longitudinal study. On the basis of the current literature we hypothesized that parents' and teachers' perceptions of children's competence would predict children's academic achievement 3 years later. Further, we assumed that

parents' and teachers' perceptions of children's competence would mediate the associations between family adversity, children's general intelligence as well as problem behavior, and children's academic performance. Thus, the study expanded on existing research by including parents' as well as teachers' perceptions of children's competence and by implementing three possible predictors of these perceptions: child behavior problems, child general intelligence, and family adversity, such as socioeconomic disadvantage and a migration background. Furthermore, the present study filled a gap in the current research by involving kindergarten to second-grade children who had not received any standardized school grades when they were first assessed; hence, their parents and teachers could not base their perceptions of the children's academic competence on grades.

Method

Sample

At the first assessment (Time 1), participants were 402 children (191 girls, 211 boys) and their parents and teachers who lived in urban and suburban regions in Austria ($n = 138$), Germany ($n = 58$), and the German-speaking part of Switzerland ($n = 206$). On all measures, no differences were revealed between the different national contexts. Children ranged in age from 5 to 7 years ($M = 6.22$ years, $SD = 0.65$ years). Sixteen percent of mothers had completed university, the majority had completed formal job training (66%), and 18% had completed only the years of compulsory schooling. Twenty-three percent of fathers had completed university, 64% had completed formal job training, and 13% had completed the years of compulsory schooling. At Time 1, parents filled in questionnaires on sociodemographic characteristics, the child's behavior problems, and their perceptions of their child's competence in mathematics, language, and science. Teachers filled in questionnaires on the child's behavior problems and their perceptions of the child's competence regarding the same school subjects. Assessment of general fluid intelligence was performed by trained and qualified school psychologists and advanced psychology students.

Three years later (Time 2) parents were contacted and asked to give information on their children's standardized school grades in mathematics, language, and science based on the latest school records. A total of 263 parents and their children (135 girls, 128 boys) returned the requested information to the study. Parents and children who participated only at the first assessment did not differ on any measures from parents and children who participated at both assessments. Only participants with complete data on both assessments were included in further analyses, resulting in a sample of 221 children (115 girls, 106 boys). Seventeen percent of the children had a migration background and were not speaking German as their first language. Seventy-one percent of the children were in kindergarten and 29% of the children were in first grade at Time 1. At Time 2, 12% of the children were in second grade, 64% were third graders, and 24% were fourth graders. There was only one child participating per family, and children in the study did not share common neighborhoods, schools, or teachers. Thirty-two families were from Austria, 64 families were from Germany, and 125 families were from the German-speaking part of Switzerland.

Measures

Children's general intelligence. The Intelligence and Development Scales (IDS; Grob, Meyer, & Hagmann-von Arx, 2009) were used at Time 1 to assess children's general fluid intelligence ($M = 100$, $SD = 15$). The IDS consists of 7 subscales covering visual perception, attention, memory, and reasoning. In the *visual perception* subtest, the child is expected to arrange cards printed with lines of varying lengths in a sequence. In the *selective attention* subtest, targets (ducks with specific characteristics) have to be checked in a speed test. In the *phonological working memory* subtest, numbers and letters of increasing length need to be repeated. An increasing number of geometric figures which have to be recognized in a new set of figures are presented in the *visual-spatial working memory* subtest. For the assessment of *long-term memory*, the examiner tells a story that the child has to retell after 30 min. In the *conceptual reasoning* subtest, the child is asked to add two examples from a selection of five pictures to three drawings of the same concept. Finally, to assess children's *figural reasoning*, the child is asked to arrange visually presented geometric figures with the aid of wooden rectangles and/or triangles. The reliability of the general intelligence test (Cronbach's α) was .85; the test-retest reliability after 15 months was .83 ($p < .001$). Correlation with the Hamburg-Wechsler Intelligence Test for Children (HAWIK-IV; Petermann & Petermann, 2008) was .83 ($p < .001$). Psychometric properties of the IDS in terms of consistency, retest reliability (after 15 months), and validity (with clinical samples) meet state-of-the art criteria (e.g., Grob et al., 2009; Hagmann-von Arx, Meyer, & Grob, 2008). For detailed information concerning the theoretical and historical background of each subtest, see Grob et al. (2009).

Family adversity. At Time 1, we included three indicators of family-related risks measured with sociodemographic questionnaires. Mothers and fathers reported their own highest level of education (years of compulsory schooling, formal job training, or university), and whether the family had a migration background (0 = *no*, 1 = *yes*). Parental education was reverse-coded so that higher scores indicated lower educational level or more educational risk. We selected the risk factors based on theoretical perspectives and prior research (e.g., Burchinal, Peisner-Feinberg, Pianta, & Howes, 2002). Because of the short window during which family-related risk factors were assessed, only time-invariant indicators of risk were included in the analyses. The reliability (Cronbach's α) of family adversity was .78.

Children's problem behavior. Children's externalizing problem behavior was assessed at Time 1 with the Externalizing Problems subscale of the Child Behavior Checklist (CBCL; Achenbach, 1992). These items reflect manifestations of distress that are expressed in outward behavior, such as persistent disobedience or physical aggression. Parents and teachers separately indicated whether each item was *often true* (3), *sometimes true* (2), or *not true* (1) of the child. We used mean scores in our analysis. Cronbach's α was .86 for parents' ratings and .82 for teachers' ratings.

Perceptions of children's competence. At Time 1, parents' and teachers' perceptions of children's competence in mathematics, language, and science were assessed, following prior research (e.g., Frome & Eccles, 1998; Jodl, Michael, Malanchuk, Eccles, & Sameroff, 2001; Pomerantz & Dong, 2006). Parents and teachers

separately rated how good they thought the child was at each of the three school subjects (mathematics, language, and science) relative to children of the same age group (1 = *at the bottom*, 5 = *at the top*). Higher numbers represent more positive perceptions. The reliability (Cronbach's α) of parents' perceptions of children's competence was .79 and of teachers' perceptions of children's competence was .72.

School performance. Three years later, at Time 2, the final outcome measure was standardized school grades in mathematics, language, and science at the end of the academic term, based on school records. Possible scores ranged from 1.0 to 6.0, with higher numbers indicative of better grades. We chose grade point average (GPA) as our distal indicator of school performance because of its significance for postsecondary education (Conley, 2007) and because it is a broad indicator of academic achievement across subject matters and teachers. As such, it can be viewed as representing a less biased indicator of school performance. Furthermore, GPA is based on performance spanning a time period (i.e., semester) rather than a single point in time.

Statistical Rationale

We analyzed all relations by means of structural equation modeling (SEM) using Amos 18 (Arbuckle, 2009). Listwise deletion was used to deal with issues of missing data. To evaluate the model fit, the χ^2 exact fit test, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI) were calculated. RMSEA values lower than .08 and CFI values above .90 were considered as satisfactory model fit indices (Byrne, 2001). For comparing specific associations within SEM, the phantom model approach was applied (Ledermann & Macho, 2009; Macho & Ledermann, 2011). The phantom model approach enables one to extract, represent, and assess one specific effect of interest within a complex model by creating a separate model (phantom model; Kenny, Korchmaros, & Bolger, 2003). In this way, the SEM program computes a separate estimation of the specific effect with its confidence intervals. Significance of a specific effect can then be tested by bootstrapping confidence intervals. The bootstrap test is incorporated in Amos 18 (Arbuckle, 2009). Thereby, estimates and standard deviations of specific effects are calculated as well as a bootstrap percentile confidence interval. From the bootstrap percentile confidence interval, which is based on $J = 5,000$ randomly selected samples from the data set, we can see where 95% of the bootstrap estimates are. If the 95% confidence intervals are significantly different from zero, it means that the bootstrap estimates are statistically significant by conventional standards. For more detailed information on the phantom model approach, see Macho and Ledermann (2011).

Observed and Latent Variables

Items of the problem behavior rated by parents and teachers were aggregated to parcels (Bandalos & Finney, 2001). For the merits of using parcels in SEM, see Little, Cunningham, Shahar, and Widaman (2002). Parcels were built according to the item-to-construct balance technique (see Little et al., 2002). Specifically, for each of the two latent variables, the two items with the highest item-total correlations were set as anchors of the respective parcels, and the two items with the lowest values were then added to

the parcels in inverted order, resulting in two parcels per latent variable in the model. Parents' and teachers' ratings of children's externalizing problem behavior resulted in the children's problem behavior factor. The general intelligence factor was composed of seven indicators: visual perception, selective attention, phonological working memory, visual-spatial working memory, long-term memory, conceptual reasoning, and figural reasoning. The family adversity factor was formed by mother's and father's educational level and family migration status (0 = *no*, 1 = *yes*). Parents' as well as teachers' perceptions of children's competence comprised perceptions with regard to children's performance in mathematics, language, and science. In order to investigate children's school performance, we created a latent factor representing standardized grades in three different domains: mathematics, language, and science.

Results

Overall Model

The structural part of the model consisted of seven latent variables (general intelligence, family adversity, problem behavior rated by parents, problem behavior rated by teachers, parents' perceptions of children's competence, teachers' perceptions of children's competence, and academic performance). The indicators of the latent variables formed the measurement part of the model and were represented by the observed variables or, as in the case of problem behavior rated by parents and teachers, by aggregating items to parcels (Bandalos & Finney, 2001; Little et al., 2002). Intercorrelations and internal consistencies among the latent variables are displayed in Table 1. The structural model was constructed on the basis of established research findings, with children's general intelligence, children's problem behavior, family adversity, and parents' and teachers' perceptions of children's competence at Time 1 as predictors of children's academic performance at Time 2. Covariances were assumed between general intelligence, family adversity, problem behavior rated by parents and problem behavior rated by teachers, as well as between parents' and teachers' perceptions of children's competence in order

Table 1
Intercorrelations (and Internal Consistencies) Among the Latent Factors in the Overall Model (N = 221)

Factor	1	2	3	4	5	6	7
1. FA	(.78)						
2. PB parents	.27**	(.86)					
3. PB teacher	.15*	.24**	(.82)				
4. IQ	-.31**	-.19*	-.22**	(.85)			
5. PC parents	-.25**	-.23**	-.21**	.59**	(.79)		
6. PC teacher	-.36**	-.19*	-.15*	.45**	.67**	(.72)	
7. AP	-.25**	-.21**	-.29**	.49**	.69**	.44**	(.82)

Note. Correlations are reported below the diagonal. Internal consistencies of the scales (Cronbach's alpha) are reported in parentheses along the diagonal. FA = family adversity; PB = problem behavior; IQ = general intelligence; PC = perceptions of children's competence; AP = academic performance.

* $p < .05$. ** $p < .01$.

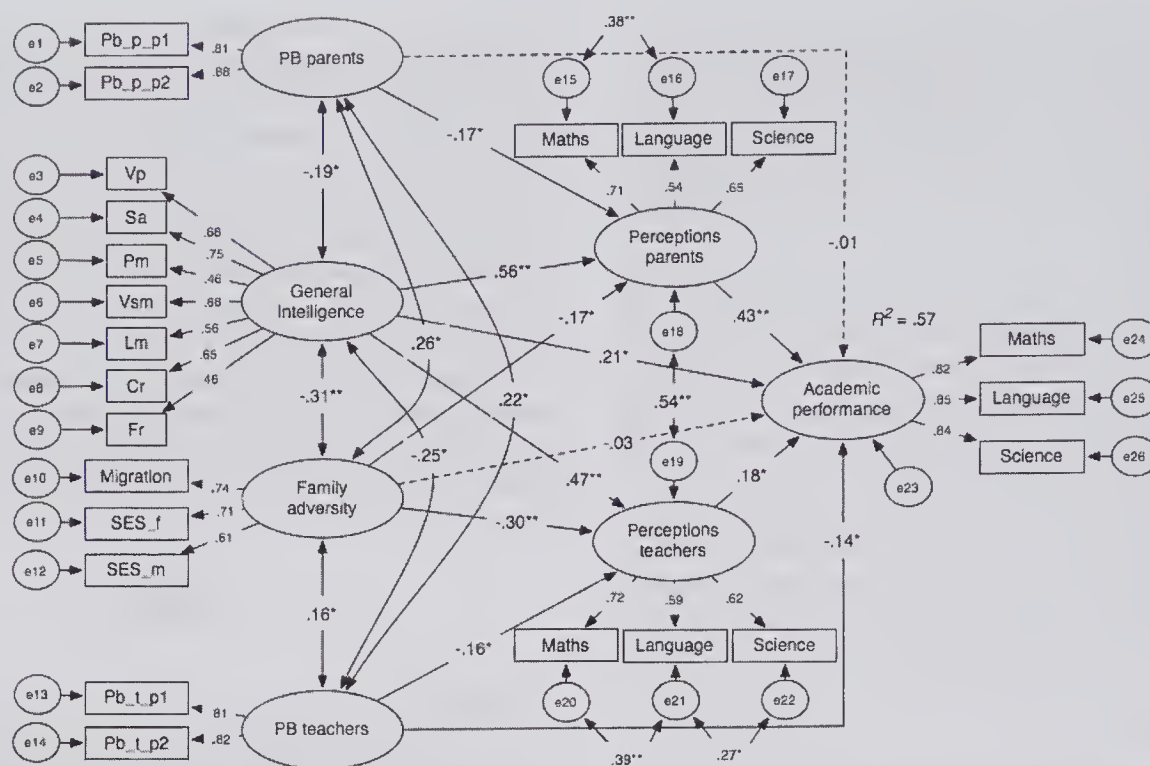


Figure 1. Overall structural equation model for the interplay among child and family characteristics, parents' and teachers' perceptions of children's competence, and children's academic performance 3 years later. Standardized coefficients are displayed. PB = problem behavior; Vp = visual perception; Sa = selective attention; Pm = phonological working memory; Vsm = visual-spatial working memory; Lm = long-term memory; Cr = conceptual reasoning; Fr = figural reasoning; SES = socioeconomic status. * $p < .05$. ** $p < .01$.

to control for shared variance. Figure 1 displays the overall model with the standardized estimate measures.

In the overall model, parents' and teachers' perceptions of children's competence were predictive of children's academic performance 3 years later ($\beta = .43, p < .01$, and $\beta = .18, p < .05$, respectively), as were children's general intelligence ($\beta = .21, p < .05$) and teachers' ratings of children's problem behavior ($\beta = -.14, p < .05$). Parents' ratings of children's problem behavior as well as family adversity were not directly predictive of children's subsequent academic performance ($\beta = -.01, p = .92$, and $\beta = -.03, p = .79$, respectively). Concerning possible predictors of perceptions of children's competence, parents' as well as teachers' perceptions were predicted by family adversity ($\beta = -.17, p < .05$, and $\beta = -.30, p < .01$, respectively), children's general intelligence ($\beta = -.56, p < .01$, and $\beta = -.47, p < .01$, respectively), and problem behavior ($\beta = -.17, p < .05$, and $\beta = -.16, p < .05$, respectively). Further, family and child characteristics were associated with each other. Family adversity was related to children's problem behavior rated by parents ($r = .26, p < .05$) and teachers ($r = .16, p < .05$) and to children's general intelligence ($r = -.31, p < .01$). Children's general intelligence was related to children's problem behavior rated by parents ($r = -.19, p < .05$) and teachers ($r = -.25, p < .05$). Children's problem behavior rated by parents and children's problem behavior rated by teachers were associated with each other ($r = .22, p < .05$). The correlation between the error terms of perceptions of children's competence in parents and teachers was significant ($r = .54, p < .01$).

Based on the suggested modification indices, three additional covariances were added to the measurement model between the

error terms of parents' perceptions of children's competences in math and language as well as between the error terms of teachers' perceptions of children's competences in math, language, and science (see Figure 1). Additional paths between parents' ratings of children's problem behavior and teachers' perceptions of children's competences as well as between teachers' ratings of children's problem behavior and parents' perceptions of children's competences resulted in a deterioration of the model and were therefore restricted to zero, $\Delta\chi^2(2) = 10.21, p < .01$. These model specifications resulted in a good representation of the data, $\chi^2(230) = 328.45, p < .001$, RMSEA = .040, CFI = .967, and accounted for 57% of the variance in children's academic performance at Time 2 (i.e., 3 years later). In order to simplify the model, the two aforementioned nonsignificant path coefficients were restricted to zero (see dashed paths in Figure 1). These simplifications did not result in a deterioration of the model fit, $\Delta\chi^2(2) = 1.03, p = .47$, and the fit indices turned out to be as good as in the overall model (RMSEA = .039, CFI = .971).

Mediation Analyses

Significance levels of comparisons between direct and indirect effects are shown in Table 2. Formal bootstrap tests of indirect effects confirmed that parents' perceptions of children's competence mediated the association between problem behavior rated by parents at Time 1 and academic performance 3 years later (bootstrap 95% CI $[-0.83, -0.17], p < .01$), as well as the association between family adversity at Time 1 and academic performance 3 years later (bootstrap 95% CI $[-0.19, -0.04], p < .05$). Furthermore, parents' perceptions of children's competence partially me-

Table 2

Standardized and Unstandardized Estimates and Confidence Interval Limits for Tests of Mediation (N = 221)

Path	Total effect			Direct effect			Indirect effect			95% CI
	<i>b</i>	(SE)	β	<i>b</i>	(SE)	β	<i>b</i>	(SE)	β	
PB parents → PC parents → AP	−0.32*	(0.14)	−.08	−0.04	(0.14)	−.01	−0.28**	(0.05)	−.07	[−0.83, −0.17]
PB teacher → PC teacher → AP	−0.30**	(0.08)	−.17	−0.21*	(0.08)	−.14	−0.09*	(0.04)	−.03	[−0.42, −0.07]
FA → PC parents → AP	−0.21*	(0.15)	−.10	−0.08	(0.15)	−.03	−0.13*	(0.06)	−.07	[−0.19, −0.04]
FA → PC teacher → AP	−0.18	(0.06)	−.08	−0.08	(0.06)	−.03	−0.10	(0.08)	−.05	[−1.12, 2.58]
IQ → PC parent → AP	1.12***	(0.53)	.45	0.63**	(0.53)	.21	0.49*	(0.12)	.24	[0.10, 0.28]
IQ → PC teacher → AP	0.96**	(0.53)	.29	0.63**	(0.53)	.21	0.32	(1.12)	.08	[−0.52, 1.88]

Note. PB = problem behavior; PC = perceptions of children's competence; AP = academic performance; FA = family adversity; IQ = general intelligence; CI = confidence interval. Bootstrap $J = 5,000$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

diated the association between children's general intelligence measured at Time 1 and children's academic performance 3 years later (bootstrap 95% CI [0.10, 0.28], $p < .05$). On the other hand, teachers' perceptions of children's competence partially mediated the association between problem behavior rated by teachers at Time 1 and academic performance 3 years later (bootstrap 95% CI [−0.42, −0.07], $p < .05$).

Discussion

The findings of the present study are consistent with prior research indicating that parents' and teachers' perceptions of children's academic competence are important to children's subsequent academic performance, even after taking into account children's general intelligence (e.g., Eccles, Wigfield, & Schiefele, 1998; Pomerantz & Dong, 2006). Moreover, the results indicated that perceptions of children's competence assessed from the perspective of informants in two different contexts, that is, parents and kindergarten teachers, were highly correlated.

Based on these long-term significant effects, possible predictors of parents' and teachers' perceptions were investigated. The present study combined child and family characteristics and showed that both of them were predictive of parents' and teachers' perceptions of children's competence. The higher children's general intelligence was, the higher parents' and teachers' perceptions of children's competence were, and the higher family adversity (in terms of lower SES and a migration background) was, as well as the higher children's behavior problems rated by parents and teachers were, the lower parents' and teachers' perceptions of children's competence were. Speaking in terms of risk factors, children with lower general intelligence scores and higher problem behavior scores from families faced with higher adversity were at greater risk of lower perceptions of competence in parents as well as teachers, with severe consequences regarding their subsequent academic performance. Furthermore, these child- and family-related risk factors were correlated with each other, pointing to the presence of risk factor aggregations. Therefore, the present study confirms the necessity of a contextualized view of children's academic achievement, which involves the simultaneous influences of child variables, family variables, as well as other exogenous variables (e.g., Bronfenbrenner & Morris, 2006).

With respect to dynamic processes, we investigated whether perceptions of children's competence mediated the associations between child as well as family characteristics and children's

academic performance 3 years later. Results demonstrated that the linkage between higher amounts of problem behavior and lower subsequent academic performance could be explained by lower perceptions of children's competence in parents and teachers. Further, parents' perceptions of children's competence fully mediated the association between family adversity and academic performance as well as partially mediated the association between children's general intelligence and academic performance. Additionally, parents' perceptions of children's competence were a stronger predictor of children's later academic performance than were teachers' perceptions of children's competence, which is in line with previous findings (e.g., Entwisle, 1997; Wigfield, Eccles, Yoon, & Harold, 1997). A possible explanation might be that teachers changed in the transition from kindergarten at Time 1 to primary school at Time 2. So the home and school contexts may largely differ in terms of stability. Together these findings underline the important role of parents' perceptions in relation to their children's future academic outcomes. Therefore, in terms of practical implications, teachers and social workers should be trained to professionally evaluate not only their own perceptions but most importantly the perceptions parents hold regarding children's academic competencies as early as kindergarten. Further, especially for parents faced with higher family adversity and children with higher amounts of problem behavior, information and support should be provided on how to optimize and accurately communicate their perceptions to their children, together with opportunities to stimulate their children's development.

Although the current research yielded important insights in understanding how perceptions of children's competence mediate the association between child and family characteristics and subsequent academic performance, there are some limitations of the study that need to be noted. First, the sample was limited in terms of the cultural background of the participating families. However, the percentage of families with a migration background in the present study was representative of the three national contexts (e.g., Lanzieri, 2008). The sample was also limited in that we did not include separate ratings of mothers' and fathers' perceptions of children's academic competence and we did not assess any child ratings. Furthermore, additional intervening variables that may explain the effect of perceptions of children's competencies on later academic outcomes should be included in future studies to specify the mediating processes (e.g., differential treatments, self-efficacy beliefs, accuracy of parental perceptions). Despite the

privileged 3-year longitudinal design of the present study, another limitation is that child and family characteristics were measured only at the first assessment. However, relying on theory and research on developmental stability, we assumed children's general intelligence and problem behavior to be relatively stable across the 3-year time span (e.g., Brame, Nagin, & Tremblay, 2001; Côté, Vaillancourt, LeBlanc, Nagin, & Tremblay, 2006; Grob et al., 2009). Further, only time-invariant indicators of family adversity were included in the study (parental educational status and migration background).

To sum up, this study responded to the repeated claim that research on perceptions of children's competence and academic performance should not only focus on parental information and high school students but should also include further child as well as family characteristics in a younger age group and in a longitudinal design. The results suggest that parents and teachers, especially parents and teachers of children with higher amounts of problem behavior and from families with higher adversity, should be supported and trained in reflecting on their own perceptions of children's academic competence in order to prevent the negative effects that lower perceptions may have on children's subsequent academic performance.

References

- Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist and 1992 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Alexander, K. L., Entwisle, D. R., & Kabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record*, 103, 760–822. doi:10.1111/0161-4681.00134
- Alexander, K. L., Entwisle, D. R., & Thompson, M. S. (1987). School performance, status relations, and the structure of sentiment: Bringing the teacher back in. *American Sociological Review*, 52, 665–682. doi:10.2307/2095602
- Arbuckle, J. L. (2009). *Amos 18 user's guide*. Chicago, IL: Amos Development Corporation.
- Ayoub, C., O'Connor, E., Rappolt, G., Vallotton, C., Raikes, H., & Chazan, R. (2009). Cognitive skills performance among young children living in poverty: Risk, change, and the promotive effect of Early Head Start. *Early Childhood Research Quarterly*, 24, 289–305. doi:10.1016/j.jecresq.2009.04.001
- Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah, NJ: Erlbaum.
- Brame, B., Nagin, D. S., & Tremblay, R. E. (2001). Developmental trajectories of physical aggression from school entry to late adolescence. *Journal of Child Psychology and Psychiatry*, 42, 503–512. doi:10.1111/1469-7610.00744
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.
- Bronfenbrenner, U., & Morris, P. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 795–828). Hoboken, NJ: Wiley.
- Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology*, 40, 415–436. doi:10.1016/S0022-4405(02)00107-3
- Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Capps, R., Fix, M., Ost, J., Reardon, J., & Passel, J. S. (2004). *The health and well-being of young children of immigrants*. Washington, DC: Urban Institute.
- Chan, D. W. (2009). Dimensionality and typology of perfectionism: The use of the Frost Multidimensional Perfectionism Scale with Chinese gifted students in Hong Kong. *Gifted Child Quarterly*, 53, 174–187. doi:10.1177/0016986209334963
- Conley, D. T. (2007). *Toward a comprehensive conception of college readiness*. Eugene, OR: Educational Policy Improvement Center.
- Côté, S. M., Vaillancourt, T., LeBlanc, J. C., Nagin, D. S., & Tremblay, R. E. (2006). The development of physical aggression from toddlerhood to pre-adolescence: A nation wide longitudinal study of Canadian children. *Journal of Abnormal Child Psychology*, 34, 71–85. doi:10.1007/s10802-005-9001-z
- Dandy, J., & Nettelbeck, T. (2002). Research note: A cross-cultural study of parents' academic standards and educational aspirations for their children. *Educational Psychology*, 22, 621–627. doi:10.1080/0144341022000023662
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35, 13–21. doi:10.1016/j.intell.2006.02.001
- Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed. In N. Eisenberg (Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 1017–1095). New York, NY: Wiley.
- Elder, G. H., Eccles, J. S., Ardelt, M., & Lord, S. (1995). Inner-city parents under economic pressure: Perspectives on the strategies of parenting. *Journal of Marriage and the Family*, 57, 771–784. doi:10.2307/353931
- Entwisle, D. R. (1997). *Children, schools, and inequality*. Boulder, CO: Westview Press.
- Friedman, T. L. (2005). *The world is flat: A brief history of the twenty-first century*. New York, NY: Farrar, Straus and Giroux.
- Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology*, 74, 435–452. doi:10.1037/0022-3514.74.2.435
- Grob, A., Meyer, C. S., & Hagmann-von Arx, P. (2009). *Intelligence and development scales (IDS)*. *Intelligenz- und Entwicklungsskalen für Kinder im Alter von 5 bis 10 Jahren*. Bern, Switzerland: Huber.
- Gutman, L. M., Sameroff, A. J., & Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: Effects of multiple social risk factors and preschool child factors. *Developmental Psychology*, 39, 777–790. doi:10.1037/0012-1649.39.4.777
- Hagmann-von Arx, P., Meyer, C. S., & Grob, A. (2008). Assessing intellectual giftedness with the WISC-IV and the IDS. *Journal of Psychology*, 216, 173–180.
- Harackiewicz, J., Barron, K., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, 94, 562–575. doi:10.1037/0022-0663.94.3.562
- Jodl, K. M., Michael, A., Malanchuk, O., Eccles, J. S., & Sameroff, A. J. (2001). Parents' role in shaping early adolescents' occupational aspirations. *Child Development*, 72, 1247–1266. doi:10.1111/1467-8624.00345
- Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3, 1–24. doi:10.1016/0022-1031(67)90034-0
- Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9, 131–155. doi:10.1207/s15327957pspr0902_3
- Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8, 115–128. doi:10.1037/1082-989X.8.2.115

- Kornilova, T. V., Kornilov, S. A., & Chumakova, M. A. (2009). Subjective evaluations of intelligence and academic self-concept predict academic achievement: Evidence from a selective student population. *Learning and Individual Differences, 19*, 596–608. doi:10.1016/j.lindif.2009.08.001
- Kuncel, N. R., Hezlett, S. A., & Ones, D. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148–161. doi:10.1037/0022-3514.86.1.148
- Kyttälä, M., & Lehto, J. E. (2008). Some factors underlying mathematical performance: The role of visuospatial working memory and non-verbal intelligence. *European Journal of Psychology of Education, 23*, 77–94. doi:10.1007/BF03173141
- Lanzieri, G. (2008). Population in Europe 2007: First results. *Eurostat Statistics in Focus, 81*, 1–12.
- Ledermann, T., & Macho, S. (2009). *Assessing mediation in simple and complex models*. Manuscript submitted for publication.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173. doi:10.1207/S15328007SEM0902_1
- Macho, S., & Ledermann, T. (2011). Estimating, testing, and comparing specific effects in structural equation models: The phantom model approach. *Psychological Methods, 16*, 34–43. doi:10.1037/a0021763
- McLeod, J. D., & Kaiser, K. (2004). Childhood emotional and behavioral problems and educational attainment. *American Sociological Review, 69*, 636–658. doi:10.1177/000312240406900502
- McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist, 53*, 185–204. doi:10.1037/0003-066X.53.2.185
- Mistry, R. S., White, E. S., Benner, A. D., & Huynh, V. W. (2009). A longitudinal study of the simultaneous influence of mothers' and teachers' educational expectations on low-income youth's academic achievement. *Journal of Youth and Adolescence, 38*, 826–838. doi:10.1007/s10964-008-9300-0
- Peet, S. H., Powell, D. R., & O'Donnel, B. K. (1997). Mother–teacher congruence in perceptions of the child's competence and school engagement: Link to academic achievement. *Journal of Applied Developmental Psychology, 18*, 373–393. doi:10.1016/S0193-3973(97)80006-8
- Petermann, F., & Petermann, U. (2008). *Hamburg-Wechsler-Intelligenztest für Kinder IV*. Bern, Switzerland: Huber.
- Phillipson, S. N. (2009). *Role of parents in children's academic achievement: A specific sociocultural context*. Köln, Germany: LAP LAMBERT Academic Publishing.
- Phillipson, S. (2010). Modeling parental role in academic achievement: Comparing high-ability to low- and average-ability students. *Talent Development & Excellence, 2*, 83–103.
- Phillipson, S., & Phillipson, S. N. (2007). Academic expectations, belief of ability, and involvement by parents as predictors of child achievement: A cross-cultural comparison. *Educational Psychology, 27*, 329–348. doi:10.1080/01443410601104130
- Pomerantz, E. M., & Dong, W. (2006). Effects of mothers' perceptions of children's competence: The moderating role of mothers' theories of competence. *Developmental Psychology, 42*, 950–961. doi:10.1037/0012-1649.42.5.950
- Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster, Germany: Waxmann.
- Rutchick, A. M., Smyth, J. M., Lopoo, L. M., & Dusek, J. B. (2009). Great expectations: The biasing effects of reported child behavior problems on educational expectancies and subsequent academic achievement. *Journal of Social and Clinical Psychology, 28*, 392–413. doi:10.1521/jscp.2009.28.3.392
- Sektnan, M., McClelland, M. M., Acock, A., & Morrison, F. J. (2010). Relations between early family risk, children's behavioral regulation, and academic achievement. *Early Childhood Research Quarterly, 25*, 464–479. doi:10.1016/j.ecresq.2010.02.005
- Taub, G. E., Keith, T. Z., Floyd, R. G., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly, 23*, 187–198. doi:10.1037/1045-3830.23.2.187
- Wigfield, A., Eccles, J. S., Yoon, K. S., & Harold, R. D. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology, 89*, 451–469. doi:10.1037/0022-0663.89.3.451
- Yamamoto, Y., & Holloway, S. D. (2010). Parental expectations and children's academic performance in sociocultural context. *Educational Psychology Review, 22*, 189–214. doi:10.1007/s10648-010-9121-z

Received May 11, 2011

Revision received October 18, 2012

Accepted November 12, 2012 ■

The Effects of Single-Sex Compared With Coeducational Schooling on Mathematics and Science Achievement: Data From Korea

Erin Pahlke
Whitman College

Janet Shibley Hyde and Janet E. Mertz
University of Wisconsin—Madison

Some U.S. school districts are experimenting with single-sex schooling, hoping that it will yield better academic outcomes for students. Empirical research on the effects of single-sex schooling, however, has been equivocal, with various studies finding benefits, disadvantages, or no effect. Most of this research is marred because families generally choose whether the child attends a single-sex or coeducational school, so there are selection effects that create pre-existing differences between students in the different types of schools. The research reported here capitalized on the case of Korea, where students are randomly assigned to single-sex or coeducational schools. Using 2007 Trends in International Mathematics and Science Study (TIMSS) data, we applied hierarchical linear modeling to account for the nesting of students within schools. Results for eighth graders indicated no differences between students in single-sex and coeducational schools in mathematics and science achievement. Results from the 2003 TIMSS data replicated the finding: students' mathematics and science achievement was unrelated to the gender composition of their school. These results call into question whether single-sex schooling has the academic advantages claimed by its proponents.

Keywords: single-sex schooling, coeducation, mathematics performance, science achievement, Korea

Supplemental materials: <http://dx.doi.org/10.1037/a0031857.supp>

In 2006, the U.S. Congress made changes to the No Child Left Behind Act of 2001 that eased restrictions on sex-segregated education in public schools. The act approved federal funding for innovative education programs, including single-sex programs within existing coeducational schools. Public schools across the country have responded by implementing single-sex education programs; based on follow-up analyses of Office of Civil Rights 2010 data, the Feminist Majority Foundation (2011) estimated that thousands of U.S. public schools offered single-sex academic classes during the 2009–2010 school year.

One of the main reasons single-sex schooling has become more popular in the United States is educators' concerns about American children's performance in mathematics and science. In recent international assessments, including Program for International Student Assessment (PISA) 2009 (Organization for Economic Cooperation and Development [OECD], 2010) and Trends in Interna-

tional Mathematics and Science Study (TIMSS) 2007 (Gonzales et al., 2008), U.S. students have lagged behind many of their international peers. This middling performance has led many observers to conclude that American education is in a state of crisis. In such a state of emergency, policy makers tend to try anything new, and single-sex education has leaped into the breach. Thus, it is critical that researchers muster the best available scientific data to investigate the effect of single-sex education on students' academic achievement, particularly in the areas of mathematics and science, to determine whether moving toward single-sex schooling will result in higher achievement for American students.

Supporters of single-sex education believe that separating boys and girls increases students' achievement and academic interest, particularly in the fields of mathematics and science and particularly for girls (James, 2009; Kessels & Hannover, 2008; Sax, 2005; Shapka & Keating, 2003). They draw on a number of arguments to support their claims about the efficacy of single-sex schooling, including (a) views that gender differences in psychological characteristics relevant to learning are substantial and biological and (b) social psychological or "girl power" approaches that highlight the negative effects of sexism in coeducational classrooms. With regard to the biological difference perspective, supporters of single-sex education argue that boys and girls do better when they receive instruction that is targeted toward supposed substantial, biologically based differences between boys and girls (Gurian, Henley, & Trueman, 2001; Sax, 2005). Supporters of this perspective argue, for example, that girls learn more when the instruction is cooperation-based, whereas boys flourish in competition-based learning environments. Supporters also claim that research suggests that girls have better hearing abilities than boys; teachers,

This article was published Online First March 18, 2013.

Erin Pahlke, Psychology Department, Whitman College; Janet Shibley Hyde, Psychology Department, University of Wisconsin—Madison; and Janet E. Mertz, McArdle Laboratory for Cancer Research, University of Wisconsin—Madison.

This article was funded in part by National Science Foundation Grant DRL-1138114. Any opinions expressed here are those of the authors and not the National Science Foundation. We thank Carey Cooper, Patrick Gonzales, and Eunjung Kim for their expert guidance and comments on drafts.

Correspondence concerning this article should be addressed to Erin Pahlke, Psychology Department, Whitman College, Walla Walla, WA 99362. E-mail: pahlke@whitman.edu

they argue, can improve student outcomes by talking more loudly to students in all-male classrooms than all-female classrooms (Sax, 2010). Related to this perspective, thousands of teachers have attended trainings through the Gurian Institute and the National Association for Single-Sex Education to learn how to teach to boys' and girls' supposed natural ways of learning (Gurian, Stevens, & Daniels, 2009).

Drawing from a social psychological approach, other supporters of single-sex schooling hold a "girl power" view and cite the problem of domineering boys in coeducational classrooms as a reason for separating the genders. In coeducational classrooms, boys tend to seek out and receive the majority of teachers' attention, particularly in math and science (Lee, Marks, & Byrd, 1994). Furthermore, some educators worry that boys' sexist attitudes and behaviors decrease girls' interest in traditionally masculine science, technology, engineering, and mathematics (STEM) fields (Sadker & Sadker, 1994). Classrooms that do not include boys, they argue, are more supportive of girls' academic achievement in counterstereotypic domains (Shapka & Keating, 2003). The reasoning goes that in single-sex classrooms, girls can develop self-confidence in mathematics and science. This view is consistent with social psychologists' emphasis on the crucial importance of social context and social interaction in influencing students' behavior (Rudman & Glick, 2008).

Opponents of single-sex schooling, in comparison, often draw on gender similarities and social justice perspectives. In response to proponents' arguments that single-sex schools are needed to address gender differences in education-relevant traits, opponents argue that the differences are actually quite small or nonexistent and that the distributions for males and females are highly overlapping (American Civil Liberties Union, 2012; Eliot, 2009; Hyde, 2005). Thus, opponents of single-sex schooling worry that gender-targeted classrooms do not address the needs of the majority of students, who are in the area of gender-overlap on the relevant traits. Furthermore, relatively little is known about differences between the brains of boys and girls, and so opponents of single-sex schooling argue that even if there are differences, there is not sufficient evidence of their relevance to classroom learning (Halpern et al., 2011).

Other opponents of single-sex schooling take a social justice perspective and argue that gender segregation is harmful in the same ways that race segregation and social class segregation are harmful. These opponents believe that separate is rarely equal and, furthermore, that children learn more when they are exposed to diverse environments that promote tolerance and cooperation (Rustad & Woods, 2004). Reducing cross-group contact in the classroom, they argue, results in boys and girls who are more gender stereotyped and who miss out on opportunities to learn from and cooperate with the other gender (Jackson & Smith, 2000). Research on preschool children's exposure to same- and other-sex peers provides support for this perspective; for example, boys with more exposure to same-sex peers become more aggressive over time (Martin & Fabes, 2001). Moreover, this effect is dose-dependent, and so the more time that boys spend with other boys, the more aggressive they become.

Hundreds of empirical studies have investigated the effects of single-sex schooling (Bracey, 2006). Unfortunately, however, the results of the studies are inconsistent and difficult to interpret. In a 2005 report commissioned by the U.S. Department of Education,

Mael, Alonso, Gibson, Rogers, and Smith found a greater number of studies demonstrating positive than negative effects of single-sex relative to coeducational schooling. However, the authors noted that nearly equal numbers of studies report mixed or no effects and positive effects of single-sex education.

It is important to note that reviewers of the single-sex schooling literature (e.g., Bracey, 2006; Mael et al., 2005) have decried the methodological weaknesses that characterize many of the studies. These methodological weaknesses make it difficult to interpret the findings. In the United States, single-sex schools are schools of choice; students (and their parents) can choose whether they attend a single-sex or coeducational school. As such, students at single-sex schools often differ in substantial ways from their peers at comparison coeducational schools, introducing potentially serious confounding variables. Prior work suggests, for example, that students at single-sex and coeducational U.S. schools differ in terms of their family resources, academic achievement, and attitudes about school before they start the single-sex or comparison program (e.g., Hayes, Pahlke, & Bigler, 2011; Marsh, 1989; Pahlke, Bigler, & Patterson, 2012). Single-sex and coeducational U.S. schools also frequently differ; for example, nearly all single-sex schools (whether private or public charter) employ some form of selective admission. As a result of these student and school selection factors, it is difficult to disentangle whether differences between the performance of students at single-sex and coeducational schools are due to the gender of the student body or to other student and school characteristics (Bracey, 2006; Mael et al., 2005).

In South Korea (hereafter, Korea), single-sex schools are not schools of choice. The city or provincial office of education in each district decides how many of their schools should be coeducational versus single-sex. In most areas, officials opt to have a mix of schooling types (T. Kim, Lee, & Lee, 2003). Students are then randomly assigned to these schools through the equalization policy. This policy, which was implemented in the 1970s, requires schools to take the students assigned to them by a district-wide lottery run by the Ministry of Education. As of 2000, all middle schools in Korea were covered by this policy (S. Kim & Lee, 2003). Compliance with the policy has been extremely high at the middle school level, although lower at the high school level, particularly in rural areas (S. Kim & Lee, 2003). The most recent reports indicate that beginning in 2011, several cities began taking parental choice into account to some extent (E. Kim, personal communication, March 16, 2012). Therefore, data collected prior to 2011 offer the best opportunity to examine the effects of school type when assignment was done randomly.

Korea offers an ideal opportunity to examine the effects of single-sex schooling without concerns about student or school selection effects. However, we are aware of only two studies that have taken advantage of the opportunity that the Korean system presents. D. H. Kim and Law (2012) investigated the effects of schooling type on 15-year-old students' mathematics achievement using data from the international study PISA. Once individual and school characteristics were controlled for, girls in single-sex schools earned slightly higher (14 points on a test with $M = 500$ and $SD = 100$) mathematics scores than girls in coeducational schools; and boys in single-sex schools earned higher mathematics scores (21 points) than boys in coeducational schools. These results suggest that students in single-sex schools in Korea perform

better in terms of mathematics than their peers at coeducational schools. However, D. H. Kim and Law (2012) did not provide information about the equivalence between the coeducational and single-sex schools, making it difficult to identify the factors that may have been underlying the difference between the single-sex and coeducational students' achievement. Moreover, some of the students in the Kim and Law sample were in high school, and random assignment appears to have been practiced less carefully at the high-school level than at the middle-school level (S. Kim & Lee, 2003). In the current study, we addressed this issue by examining differences among the student, teacher, and school characteristics associated with the single-sex and coeducational schools. We also extended the D. H. Kim and Law (2012) findings by examining achievement in both mathematics and science among a younger, middle-school cohort.

Park and Behrman (2010) completed the only other study we are aware of that examined the effects of school type on Korean students' outcomes. In their working paper, the researchers examined the effects of attending a single-sex or coeducational high school on the college attendance rates of students in Seoul. Results suggested that students who attend single-sex schools are more likely to go on to 4-year colleges (and less likely to go to 2-year colleges) than their peers at coeducational schools. The authors checked for the equivalence of the coeducational and single-sex high schools in terms of the socioeconomic resources of the schools and individual students as well as the teacher-student ratios, and results generally supported the effectiveness of the equalization policy in Korea. Specifically, students in single-sex and coeducational schools came from families with similar levels of household income and had parents with similar levels of education and homeownership rates. There was, however, a higher percentage of male teachers at the all-boy schools than at the coeducational schools. The authors did not, however, use multilevel models. The use of multilevel models is important because they account for the nested nature of the data (i.e., students within schools). Students within the same school tend to be similar in many respects that are unrelated to the gender composition of the classroom (e.g., family resources, teacher qualifications). The use of ordinary least squares regression models that do not account for the nesting potentially results in findings that overestimate the effects of the gender composition of schools. Therefore, in the current study we used hierarchical linear modeling (HLM) to account for the nesting of students within schools.

The Current Study

Past research on the effects of single-sex compared with coeducational schooling has been marred by the fact that students (and their parents) typically exercise choice in which type of school they attend, introducing selection effects. In the current study, we took advantage of the random assignment of students within districts to single-sex or coeducational schools in Korea to examine the effects of single-sex versus coeducational schooling on students' academic achievement in a context free of selection effects. To test arguments that single-sex schooling is particularly advantageous for girls in nontraditional academic domains, we examined achievement in both mathematics and science. We examined these questions with an eighth-grade, middle-school sample, which provided a strong test of the effects

of single-sex versus coeducational schooling since the students had been in the school context for both seventh and eighth grades at the time of assessment. We also examined these questions with two different sets of the data—collected from eighth graders in both 2007 and 2003—as a test of the robustness and replicability of the findings. Using the TIMSS data, we applied multilevel models that accounted for the nesting of students within schools. And we considered the achievement of girls and boys separately to examine hypotheses related to the supposed differential effects of single-sex schooling on girls' compared with boys' mathematics and science performance.

Method

Data and Sample

The data for the present study came from the Korean sample of the TIMSS from 2007 and 2003. In both waves, nationally representative samples of Korean eighth graders were created with a multistage sampling frame. The sample design included explicit stratification by 16 provinces and implicit stratification by urbanization (urban, suburban, rural) and school gender composition (coeducational, all girls, all boys; see <http://timssandpirls.bc.edu/TIMSS2007> and <http://timssandpirls.bc.edu/TIMSS2003>).

In 2007, participants were 4,240 eighth-grade students (2,016 girls and 2,224 boys) in the Republic of Korea. In all, 150 Korean schools participated in TIMSS 2007. These schools were coeducational (56.7%), all-boys (22.6%), and all-girls (20.7%) schools. One eighth-grade mathematics and one eighth-grade science class were sampled from each participating school.

In 2003, participants were 5,309 eighth-grade students (2,533 girls and 2,776 boys) in the Republic of Korea. In all, 149 schools participated in TIMSS 2003. These schools were coeducational (46.3%), all-boys (28.2%), and all-girls (25.5%) schools. One eighth-grade mathematics and one eighth-grade science class were sampled from each participating school.

Data were collected in the fall of 2007 and 2003 and included all eighth-grade students in the targeted mathematics and science classes. Students completed measures of mathematics and science achievement as well as an extensive background questionnaire. Teachers in the targeted classrooms also completed surveys about their own background and the classroom environment. School administrators completed surveys on the physical, organizational, and fiscal characteristics of their schools, as well as the school's learning environment and programs.

To maximize the use of available information and minimize bias, we imputed missing data using the multiple imputation (MI) procedure in SAS (note that there were no missing data on the achievement tests). This procedure uses information from available data to create 10 data sets in which missing values are replaced with different sets of equally plausible values. Results are generated from each data set and then combined to create valid statistical inferences for the parameters of interest.

Measures

Achievement. TIMSS included assessments of students' mathematics and science achievement. These assessments examine topics that students are expected to have learned (content domains)

and skills students are expected to have developed (cognitive domains). In TIMSS 2007 math, the content domains are number, algebra, geometry, and data and chance; in TIMSS 2007 science, the content domains are biology, chemistry, physics, and earth science. In both TIMSS 2007 mathematics and science, the cognitive domains are knowing, applying, and reasoning. The TIMSS 2003 domains are slightly different (e.g., environmental science was included as a science content domain; see Mullis et al., 2003, for a complete description). Overall mathematics and science scale scores are created by averaging the relevant content and cognitive domain scores. For each measure, TIMSS creates five plausible values that are then combined in HLM to create an estimate of achievement (Olson, Martin, & Mullis, 2008). Each scale ranges from 0 to 1,000, with the mean set at 500 and the standard deviation set at 100 based upon the scores of the benchmark participants (Mullis et al., 2005).

Mother's and father's education. Students reported on both their father's and mother's level of education using a 6-point scale: 1 = *did not finish elementary school or middle school or did not go to school*, 2 = *finished middle school*, 3 = *finished high school*, 4 = *finished junior college*, 5 = *finished university*, and 6 = *finished graduate school*. These measures (and all other measures except where specifically noted) were exactly the same in TIMSS 2007 and 2003.

Family possessions. In 2007, students reported whether they had nine possessions in their homes: (a) calculator, (b) computer, (c) study desk, (d) dictionary, (e) Internet connection, (f) car, (g) air conditioner, (h) DVD or VCR player, and (i) digital camera. In 2003, students reported whether they had 16 possessions in their home: (a) calculator, (b) computer, (c) study desk, (d) dictionary, (e) own study room, (f) car, (g) air conditioner, (h) camera, (i) notebook computer, (j) audio recorder, (k) videotape recorder, (l) camcorder, (m) cell phone, (n) printer, (o) washing machine, and (p) refrigerator for kimchi. The total number of items the students reported among their family possessions served as a proxy measure for family income.

Time spent on math and science homework. In two separate questions, students reported on the average number of minutes they usually spent on their math homework and the average number of minutes they usually spent on their science homework, using a 6-point scale: 1 = *zero minutes*, 2 = *1–15 minutes*, 3 = *16–30 minutes*, 4 = *31–60 minutes*, 5 = *61–90 minutes*, and 6 = *more than 90 minutes*.

Expected educational attainment. Students responded to the question, "How far in school do you expect to go?" using a 4-point scale: 1 = *finish high school*, 2 = *finish junior college*, 3 = *finish university*, and 4 = *finish graduate school*.

Economic disadvantage of students. School administrators responded to the question, "Approximately what percentage of students in your school come from economically disadvantaged homes?" using a 4-point scale: 1 = *0%–10%*, 2 = *11%–25%*, 3 = *26%–50%*, and 4 = *more than 50%*.

Size of community. School administrators reported the size of the community where the school is located using a 6-point scale: 1 = *3,000 or fewer people*, 2 = *between 3,001 and 15,000 people*, 3 = *between 15,001 and 50,000 people*, 4 = *between 50,001 and 100,000 people*, 5 = *between 100,001 and 500,000 people*, and 6 = *more than 500,000 people*.

Instructional time. In 2007, school administrators reported the number of instructional days per school year. In 2003, school administrators reported the number of instructional hours per school year.

School enrollment. School administrators reported the number of students in the school.

Mathematics and science instructional resources. TIMSS created two indices of school resources available for mathematics and science instruction. The indices take into account principals' reports of instructional materials (e.g., textbooks), budget for supplies, school building and grounds, heating/cooling and lighting system, instructional spaces, computers, library materials, and audio-visual resources (see Olson et al., 2008, for details). The mathematics and science indices were both reliable measures in the Korean sample (in 2007, α s = .87 and .89, respectively).

Teacher experience. Mathematics and science teachers reported the number of years they had been teaching.

Teacher education. Mathematics and science teachers reported the highest level of formal education they had completed using a 5-point scale: 1 = *did not finish high school*, 2 = *finished high school*, 3 = *finished junior college*, 4 = *finished university*, and 5 = *finished graduate school*.

Overview of Data-Analytic Strategy

Data analysis included four steps. In the first three steps, we used the TIMSS 2007 data. First, we tested for the presence of student- and school-based differences between the coeducational and single-sex samples. Korea randomly assigns students to neighborhood schools, and parents and students supposedly did not have a choice about school placement when the data were collected. Furthermore, coeducational and single-sex schools are prevalent in all types of Korean communities (e.g., urban, rural). Preexisting differences in student and school characteristics between the coeducational and single-sex samples were tested as a check of the purported random assignment. In the second step, we tested whether achievement varied as a function of school type. That is, we tested whether students' achievement on the mathematics and science achievement tests varied as a function of whether students attended a coeducational or single-sex school. These initial models assessed students' achievement on the math and science scale scores, along with the underlying content and cognitive domains, to examine the presence of overall differences due to school type. In the third step, we included school- and student-level control variables in the models. We tested the models using HLM to account for the nested nature of the data (i.e., students nested within schools). Finally, in the fourth step, we retested the final models using TIMSS 2003 data as a check of the robustness of the findings.

The intraclass correlations (ICCs) from the baseline models predicting mathematics and science achievement with no student- or school-level controls ranged from .06 to .10 in TIMSS 2007, which indicates that between 6% and 10% of the variance in achievement was between schools in the current sample (Raudenbush & Bryk, 2002). Although slightly lower than typical ICCs of .22 across grades in U.S. samples (Hedges & Hedberg, 2007), these ICCs were all significantly different from zero, thus indicating that nested models are appropriate. HLM also allowed analyses to be weighted using the appropriate TIMSS weight (TOTWGT),

which corrected for deviations from representativeness due to unequal probability of sample selection and nonrandom response bias. All models were tested separately for boys and girls to examine potential differences based on student gender.

Results

Test of Presence of Student- and School-Based Differences

One of the largest methodological issues in research on the effectiveness of single-sex schooling is the concern that differences between students' performance in coeducational and single-sex schooling may be due to student- or school-based factors that are unrelated to the gender composition of the classroom. Therefore, before examining differences between boys' and girls' mathematics and science achievement in coeducational versus single-sex classrooms, we tested for student- and school-based differences.

To examine differences between coeducational and single-sex students, we performed a series of *t* tests. Analyses were conducted separately for boys and girls. To control for Type I error, we adjusted the alpha level to 0.004 using a Bonferroni correction that took into account the number of statistical tests related to student-based differences. Table 1 presents information about differences in student characteristics. No significant differences exist between boys in the coeducational schools and boys in the single-sex schools or between girls in the coeducational schools and girls in the single-sex schools in terms of demographics, family resources, or expected educational attainment. These findings support the belief that students really were randomly assigned to single-sex or coeducational schools; there appear to be no selection effects.

To assess the magnitude of these and all other differences reported in the article, we computed measures of effect size (Table 1; Cohen's $d = M_1 - M_2/s_{\text{pooled}}$; where M_1 is the mean for coed and M_2 is the mean for single-sex). The large sample size ($N = 2,016$ girls and 2,024 boys) allows relatively small differences to

reach significance. Thus, we were interested in the effect size of the difference between the coeducational and single-sex samples. Cohen (1988) provided guidelines for the interpretation of effect sizes: effect sizes of $d = 0.20, 0.50$, and 0.80 are considered small, medium, and large, respectively. Based on these guidelines, the effect sizes associated with all student-based differences are negligible.

To examine potential differences between the coeducational and single-sex schools, we conducted a series of one-way analyses of variance, with school gender composition (coeducation, all-girls, all-boys) as the between factor. Again, a Bonferroni correction was used that resulted in an adjusted alpha level of .01. Coeducational, all-girls, and all-boys schools were similar in terms of school and community size, percentage of students classified as economically disadvantaged, teacher experience, and teacher education (descriptive statistics are presented in online supplemental tables). Teacher gender varied as a function of the school type; teachers were 65% female at coeducational schools, compared with 70% female at all-girls schools and 50% female at all-boys schools. A chi-square analysis indicated that coeducational schools had a higher percentage of female teachers than all-boys schools did, $\chi^2(1, N = 341) = 6.23, p = .014$.

Mathematics and Science Achievement

As a first test of the effect of school type (single-sex vs. coed) on students' mathematics and science performance, we examined the mean differences between students attending single-sex and coeducational schools in mathematics and science achievement with a series of *t* tests (Bonferroni corrected alpha level = .003). These means come from nested models, which account for the shared experiences of students within the same classroom, but they do not include student- or school-level controls (for unnested models, see Kane & Mertz, 2012). The means presented here reflect the overall difference between students in single-sex versus coeducational programs, without accounting for possible selection effects. We investigated the separate cognitive and content do-

Table 1
Student Characteristics in Trends in International Mathematics and Science Study (TIMSS) 2007 by Gender and School Type

Student characteristics	Girls							Boys						
	Coed (<i>n</i> = 1,139)		Single-sex (<i>n</i> = 877)		<i>t</i>	<i>p</i>	<i>d</i>	Coed (<i>n</i> = 1,267)		Single-sex (<i>n</i> = 957)		<i>t</i>	<i>p</i>	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Age (in years)	14.27	.34	14.31	0.30	−2.83	.01	−.12	14.28	0.36	14.31	0.31	−1.85	.06	−.09
Mother's education ^a	3.58	1.91	3.44	1.70	1.65	.10	.08	3.72	1.94	3.67	1.41	0.64	.52	.03
Father's education ^a	3.99	1.90	3.84	1.95	1.63	.11	.08	4.14	2.23	4.05	1.70	0.98	.33	.05
Family possessions (no.) ^b	7.63	1.69	7.44	1.78	2.43	.02	.11	7.62	2.13	7.60	1.86	0.24	.81	.01
Math homework time (min) ^c	2.88	1.25	2.91	1.69	−0.42	.68	−.02	2.91	1.62	2.89	1.96	0.23	.82	.01
Science homework time (min) ^c	2.88	1.43	2.92	2.15	0.39	.70	.02	2.95	1.55	2.88	1.01	1.09	.28	.05
Expected educational attainment ^d	3.08	0.64	3.07	0.56	0.35	.73	.02	3.07	0.66	3.06	0.58	0.35	.73	.02

Note. Means and standard deviations were computed with the JACKGEN SAS macro, which utilizes a jackknife repeated replication methodology to compute the appropriately weighted standard errors. Positive *t* and *d* statistics indicate a coeducational school advantage; negative *t* and *d* statistics indicate a single-sex school advantage.

^a Rated on a scale ranging from 1 (*did not finish high school*) to 6 (*finished graduate school*). ^b Reported the presence of nine possessions in the home (1–9). ^c Rated on a scale ranging from 1 (*zero minutes*) to 6 (*more than 90 minutes*). ^d Rated on a scale ranging from 1 (*finish high school*) to 4 (*finish graduate school*).

mains, as well as the overall scale scores, to determine whether students in one type of schooling did particularly well in any of the individual areas of assessment. As can be seen in Table 2, a consistent pattern emerged: single-sex and coeducational students (regardless of the students' gender) performed similarly in all mathematics and science domains. These results indicate that overall, students' mathematics and science performance does not differ based on schooling type in Korea.

To examine whether there were significant effects of schooling type once school- and student-level controls were accounted for, we next tested the full HLM models that included all school- and student-level background variables. The results from 2007 for the math and science scale scores, by student gender, are presented in Table 3. Results from models using the 2003 data are presented in Table 4. Models related to the specific content and cognitive domains suggest similar results and are available from the authors. It should be noted that gender composition of the school was not a significant predictor in any of the models; in other words, girls' and boys' performance in all content and cognitive mathematics and science domains did not significantly differ as a function of their attending a coeducational or single-sex school.

Instead, for both the 2007 and the 2003 data sets, students' performance was consistently significantly predicted by factors related to socioeconomic status; students (both boys and girls) performed better on the mathematics and science exams when their fathers had more education, their families had more resources, and a lower proportion of their schoolmates came from economically disadvantaged families. Both boys' and girls' mathematics performance was predicted by the

amount of time spent on homework; students do *worse* when they spend relatively more time on mathematics homework (or students spend more time on homework when they are performing poorly). Other factors (i.e., teachers' experience, instructional time, the size of communities and schools, and students' mother's education) were predictive of achievement in some domains for either boys or girls. However, these models emphasize that it is not the gender composition of the schools that determines mathematics and science performance, but rather family and school resources.

Discussion

The No Child Left Behind Act calls for "scientifically based research" to guide education practices and programs (U.S. Department of Education, 2003). Despite numerous empirical studies of the effects of single-sex schooling, conclusions have been equivocal because of the inconsistency of findings and because of selection effects introduced by families choosing which type of school the student attends. In the current research, we capitalized on the case of Korea, where students were randomly assigned to single-sex or coeducational middle schools (seventh through ninth grades) based on their residential districts in 2007 or 2003. Overall, the results indicated no differences in mathematics and science performance between eighth graders in single-sex compared with those in coeducational schools.

Analysis of student, teacher, and school variables confirmed random assignment and the absence of selection effects. No differences between school types in terms of background variables were signifi-

Table 2
Means of Scores on Scales, Content, and Cognitive Domains in Math and Science in the Trends in International Mathematics and Science Study (TIMSS) 2007 by Gender and School Type

Domain	Girls							Boys						
	Coed (n = 1,139)		Single-sex (n = 877)		t	p	d	Coed (n = 1,267)		Single-sex (n = 957)		t	p	d
	M	SD	M	SD				M	SD	M	SD			
Math scale	596.87	84.86	590.25	85.69	1.73	.08	.08	597.14	89.44	599.11	91.51	-0.51	.61	-.02
Content														
Algebra	597.94	103.20	590.48	102.21	1.62	.11	.07	594.13	106.69	595.96	109.09	-0.40	.69	-.02
Data and chance	581.8	71.09	576.56	71.21	1.64	.10	.07	577.43	75.07	580.05	76.42	-0.81	.42	-.03
Number	576.82	89.99	568.89	90.37	1.96	.05	.09	589.16	93.35	592.27	96.08	-0.77	.44	-.03
Geometry	586.23	77.64	581.39	77.49	1.39	.17	.06	585.95	81.13	588.14	82.11	-0.63	.53	-.03
Cognitive														
Knowing	598.91	84.50	591.42	85.38	1.96	.05	.09	593.85	87.86	595.75	90.22	-0.50	.62	-.02
Applying	593.49	86.52	586.99	86.13	1.68	.09	.08	594.45	89.45	599.69	92.55	-1.34	.18	-.06
Reasoning	580.33	86.45	571.62	88.13	2.22	.03	.10	578.98	90.54	580.34	93.11	-0.35	.73	-.01
Science scale	549.66	70.02	546.70	70.38	0.94	.35	.04	555.25	77.42	558.12	74.91	-0.88	.38	-.04
Content														
Chemistry	536.02	68.53	534.38	69.09	0.53	.56	.02	534.09	75.27	537.48	76.09	-1.05	.30	-.04
Earth science	530.33	61.10	528.91	60.41	0.52	.60	.02	543.34	67.71	548.15	66.12	-1.68	.09	-.07
Biology	546.17	63.92	544.87	63.46	0.45	.65	.02	548.35	71.06	549.75	69.02	-0.47	.64	-.02
Physics	564.51	75.00	563.03	74.46	0.44	.66	.02	575.04	83.70	579.51	81.21	-1.27	.21	-.05
Cognitive														
Knowing	535.35	66.14	534.42	63.97	0.32	.75	.01	548.82	71.42	550.88	68.43	-0.69	.49	-.03
Applying	544.34	67.00	542.01	66.09	0.78	.44	.04	548.16	73.71	551.09	70.46	-0.95	.34	-.04
Reasoning	558.00	67.32	554.59	67.24	1.13	.26	.05	557.09	73.48	561.81	70.47	-1.54	.13	-.07

Note. Means and standard deviations were computed using multilevel models in hierarchical linear modeling, which took into account the five plausible values of each score and the appropriate weights. Positive *t* and *d* statistics indicate a coeducational school advantage; negative *t* and *d* statistics indicate a single-sex school advantage.

Table 3

Results From Hierarchical Linear Models Predicting the Math and Science Scale Scores in the Trends in International Mathematics and Science Study (TIMSS) 2007 by Gender

Variable	Girls		Boys	
	Math <i>b</i> (SE)	Science <i>b</i> (SE)	Math <i>b</i> (SE)	Science <i>b</i> (SE)
School-level				
Gender composition of school (1 = single-sex)	-1.14 (6.03)	0.75 (4.75)	1.25 (5.89)	1.69 (5.05)
Average teacher experience	0.20 (0.28)	-0.11 (0.24)	0.39 (0.39)	0.13 (0.27)
Average teacher education	2.55 (10.20)	6.17 (7.70)	-1.48 (13.75)	-5.76 (10.92)
Percentage of female teachers	7.29 (10.82)	4.92 (8.85)	-14.94 (9.81)	-10.97 (8.01)
Economic disadvantage of students	-11.44 (2.76)***	-7.91 (2.18)***	-13.67 (3.09)***	-9.14 (2.53)***
Size of community	3.45 (3.16)	0.98 (2.26)	8.59 (4.25)*	5.43 (3.69)
Instructional time	0.15 (0.48)	0.24 (0.34)	0.13 (0.83)	0.22 (0.61)
School enrollment	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)*	0.01 (0.01)
Instructional resources ^a	0.80 (6.38)	1.03 (4.88)	-2.78 (6.65)	-0.88 (5.33)
Student-level				
Age	2.82 (6.30)	9.85 (5.58)	2.16 (6.52)	9.33 (5.68)
Mother's education	6.06 (2.48)*	3.30 (2.11)	3.68 (2.12)	1.60 (1.95)
Father's education	9.21 (2.18)***	9.64 (1.94)***	10.45 (2.22)***	9.94 (1.90)***
Family possessions	13.68 (1.68)***	8.58 (1.58)***	11.13 (1.81)***	7.48 (1.57)***
Homework time ^a	-8.18 (2.60)**	-0.08 (2.05)	-11.20 (2.69)***	-1.45 (2.19)

Note. Gender composition of the school is uncentered. School-level control variables are all centered around the overall mean. Student-level control variables are all centered around the group means.

^a Instructional resources and homework time are subject specific, and so different versions of the same items are used in the math and science models.

* $p < .05$. ** $p < .01$. *** $p < .001$.

cant. Gender of the teacher did differ between the school types, with 70% female teachers in all-girls schools, 65% female teachers in coeducational schools, and 50% female teachers in all-boys schools. However, percentage of female teachers in the school was not a significant predictor of mathematics or science performance for either girls or boys. Results from the 2003 data, which are available upon request, indicated the same patterns. Assignment to single-sex versus coeducational schools seems to be random in Korea, and that random

assignment results in schools that differ only in terms of the gender composition of the student body; therefore, these data sets provide an excellent opportunity to test the effects of single-sex schooling, free of the selection effects that are present in United States, European, and Australian data.

The gender composition of the middle schools was not related to either boys' or girls' performance in mathematics and science. In other words, boys at coeducational schools performed as well as boys

Table 4

Results From Hierarchical Linear Models Predicting the Math and Science Scale Scores in the Trends in International Mathematics and Science Study (TIMSS) 2003 by Gender

Variable	Girls		Boys	
	Math <i>b</i> (SE)	Science <i>b</i> (SE)	Math <i>b</i> (SE)	Science <i>b</i> (SE)
School-level				
Gender composition of school (1 = single-sex)	-8.62 (4.82)	-4.21 (3.63)	-7.01 (4.73)	0.38 (3.75)
Average teacher experience	1.20 (0.47)*	0.93 (0.39)*	0.81 (0.49)	0.67 (0.37)
Average teacher education	-2.74 (7.35)	1.02 (5.88)	-2.17 (9.42)	5.13 (7.57)
Percentage of female teachers	1.81 (7.67)	9.32 (6.96)	3.35 (8.82)	4.90 (7.03)
Economic disadvantage of students	-12.34 (2.63)***	-10.79 (1.92)***	-11.64 (2.31)***	-8.88 (1.98)***
Size of community	6.11 (2.49)*	2.47 (1.60)	5.15 (2.41)*	2.91 (1.64)
Instructional time	-0.07 (0.03)*	-0.05 (0.02)*	-0.05 (0.02)	-0.03 (0.02)
School enrollment	0.02 (0.01)	0.01 (0.01)	0.02 (0.01)*	0.01 (0.01)*
Instructional resources ^a	2.73 (4.08)	0.70 (3.15)	0.20 (4.18)	0.92 (3.67)
Student-level				
Age	-2.17 (6.45)	7.68 (4.54)	8.80 (5.46)	6.75 (4.89)
Mother's education	3.67 (1.36)**	3.74 (1.13)**	7.86 (1.42)***	4.01 (1.24)**
Father's education	5.58 (1.30)***	4.43 (1.34)**	5.47 (1.31)***	3.76 (1.22)**
Family possessions	8.25 (0.99)***	4.43 (0.87)***	7.86 (1.03)***	3.92 (0.90)***
Homework time ^a	-3.21 (2.81)	3.52 (1.70)*	-8.92 (2.40)***	1.51 (1.70)

Note. Gender composition of the school is uncentered. School-level control variables are all centered around the overall mean. Student-level control variables are all centered around the group means.

^a Instructional resources and homework time are subject specific, and so different versions of the same items are used in the math and science models.

* $p < .05$. ** $p < .01$. *** $p < .001$.

at all-boys schools, and girls at coeducational schools performed as well as girls at all-girls schools. Moreover, these findings were consistent regardless of whether student- and school-level controls were included in the models and whether data from 2007 or 2003 were utilized, providing additional support for the stability and accuracy of the results.

Our conclusions drawn from middle school data differ from those of two previous studies of the effects of single-sex schooling in high schools in Korea (D. H. Kim & Law, 2012; Park & Behrman, 2010). Park and Behrman did not use multilevel modeling to account for the nesting of students within schools. Therefore, it is likely that they underestimated the standard error of scores, thus inflating significance tests. Moreover, their outcome variable was attendance at a 4-year college, which is distinct from mathematics and science performance. Finally, Park and Behrman's sample included only students attending schools in Seoul, which may also have affected their results.

Regarding the second study, D. H. Kim and Law (2012) found a small advantage for 15-year-old girls and a somewhat larger advantage for 15-year-old boys attending single-sex middle and high schools on the PISA measure of mathematics performance. Kim and Law correctly used HLM to account for the nesting of students within schools. However, the random assignment of students to high schools is not as stringent as it is to middle schools in Korea. Furthermore, differences between their findings and ours may also be accounted for in part by differences between the PISA mathematics test and the TIMSS mathematics test. TIMSS testing focuses on the attained curriculum (i.e., what students have learned in the classroom), whereas PISA focuses more on applications and reasoning beyond classroom learning (Else-Quest, Hyde, & Linn, 2010).

Overall, then, differences between the findings from the current study and these previous studies may be related to different statistical methods (failure to account for nesting) or differences in the age of the students and the type of the schools (i.e., middle vs. high schools). Until these factors can be explored, educational policy makers should be cautious in using either the Park and Behrman (2010) or D. H. Kim and Law (2012) findings as cause for further implementation of single-sex schooling. Taken together, the studies using Korean data suggest that single-sex schooling does not have consistently positive effects for student achievement.

It should be noted that the results of the current study are consistent with the few other recent studies that have attempted to disentangle the effects of selection and gender composition. Jackson (2012) examined the performance of students in single-sex and coeducational schools in Trinidad and Tobago. Once student selection and preference factors were accounted for, there was no effect of schooling type on achievement. Similar findings have been reported in U.S. school samples (i.e., Hayes et al., 2011). These studies provide additional support for the argument that the positive effects of single-sex schooling that have been reported in the past are a result of student or school selection factors, rather than the gender composition of the school.

The results of the current study are also consistent with a multitude of studies that indicate that the social and economic resources of families and schools are the strongest predictors of students' achievement (e.g., Melhuish et al., 2008). In this study, father's education and family possessions, as well as the economic

disadvantage of the school, were all significant predictors of mathematics and science performance for both boys and girls.

Strengths and Limitations

A major strength of the current study is the Korean system of random assignment of students to single-sex or coeducational schools, permitting causal inferences. Such a design is nearly impossible in the United States, where students and parents have the right to choose which type of school or classroom they prefer, which introduces substantial confounding factors. Analyses of student, teacher, and school characteristics confirmed that the groups were equivalent, which highlights the validity of these data. This study is also the first we are aware of to explore the effects of single-sex versus coeducational instruction in middle schools in Korea. Advocates of single-sex schooling often argue that gender-segregated schooling is particularly beneficial for early adolescents (e.g., Novotney, 2011), and so examining data from middle-school students is especially important.

Differences in cultural values and educational system structures may limit the generalizability of these findings to the United States or other Western nations. Replication studies would be helpful. Nonetheless, this limitation must be balanced against the design strengths presented by random assignment. Finally, in this study, we examined only mathematics and science performance. In particular, we did not examine gender-stereotyped attitudes, which may respond to the sex composition of the school. Future research should explore these outcomes.

Conclusions

Past research on the effects of single-sex schooling has led to inconclusive results for several reasons, one being that typically families have a choice about which type of school the student attends, introducing preexisting differences among students in the different school types. In this research, we capitalized on the case of Korea, which practices random assignment of students to single-sex or coeducational schools. The results consistently indicated no differences between students in single-sex versus coeducational middle schools in mathematics and science performance. In conclusion, then, single-sex schooling did not yield any benefits to middle school boys or girls in terms of their mathematics and science achievement.

References

- American Civil Liberties Union. (2012). *Preliminary findings of ACLU "Teach kids, not stereotypes" campaign*. New York, NY: Author.
- Bracey, G. W. (2006). *Separate but superior? A review of issues and data bearing on single-sex education*. Tempe: Arizona State University, Education Policy Research Unit.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Eliot, L. (2009). *Pink brain, blue brain: How small differences grow into troublesome gaps, and what we can do about it*. New York, NY: Houghton Mifflin.
- Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136, 103–127.
- Feminist Majority Foundation. (2011). *State of public school sex segregation in the United States: Part 1. Patterns of K–12 single-sex public education in the U.S. (2007–10)*. Washington, DC: Author.

- Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2009-001 Revised). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
- Gurian, M., Henley, P., & Trueman, T. (2001). *Boys and girls learn differently! A guide for teachers and parents*. New York, NY: Jossey-Bass.
- Gurian, M., Stevens, K., & Daniels, P. (2009). Single-sex classrooms are succeeding. *Educational Horizons*, 87, 234-245.
- Halpern, D. F., Eliot, L., Bigler, R. S., Fabes, R. A., Hanish, L. D., Hyde, J., . . . Martin, C. L. (2011, September 23). The pseudoscience of single-sex schooling. *Science*, 333, 1706-1707. doi:10.1126/science.1205031
- Hayes, A. R., Pahlke, E., & Bigler, R. S. (2011). Testing the efficacy of single-sex education: An investigation of selection effects and school quality. *Sex Roles*, 65, 693-703. doi:10.1007/s11199-010-9903-2
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60-87. doi:10.3102/0162373707299706
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60, 581-592. doi:10.1037/0003-066X.60.6.581
- Jackson, C. K. (2012). Single-sex schools, student achievement, and course selection: Evidence from rule-based student assignments in Trinidad and Tobago. *Journal of Public Economics*, 96, 173-187. doi:10.1016/j.jpubeco.2011.09.002
- Jackson, C., & Smith, I. D. (2000). Poles apart? An exploration of single-sex and mixed-sex educational environments in Australia and England. *Educational Studies*, 26, 409-422. doi:10.1080/03055690020003610
- James, A. N. (2009). *Teaching the female brain: How girls learn math and science*. Thousand Oaks, CA: Corwin Press.
- Kane, J. M., & Mertz, J. E. (2012). Debunking myths about gender and mathematics performance. *Notices of the AMS*, 59, 10-21. doi:10.1090/noti790
- Kessels, U., & Hannover, B. (2008). When being a girl matters less: Accessibility of gender-related self-knowledge in single-sex and coeducational classes and its impact on students' physics-related self-concept of ability. *British Journal of Educational Psychology*, 78, 273-289. doi:10.1348/000709907X215938
- Kim, D. H., & Law, H. (2012). Gender gap in maths test scores in South Korea and Hong Kong: Role of family background and single-sex schooling. *International Journal of Educational Development*, 32, 92-103. doi:10.1016/j.ijedudev.2011.02.009
- Kim, S., & Lee, J.-H. (2003). *The secondary school equalization policy in South Korea* (KDI School Working Paper No. 02-05). Retrieved from <https://pantherfile.uwm.edu/kim/www/papers/Equalization5.doc>
- Kim, T., Lee, J.-H., & Lee, Y. (2003). Mixing versus sorting in schooling: Evidence from the equalization policy in South Korea (KDI Working Paper 03-07). Retrieved from <http://ssrn.com/abstract=482962>
- Lee, V. E., Marks, H. M., & Byrd, T. (1994). Sexism in single-sex and co-educational independent secondary school classrooms. *Sociology of Education*, 67, 92-120. doi:10.2307/2112699
- Mael, F., Alonso, A., Gibson, D., Rogers, K., & Smith, M. (2005). *Single-sex versus coeducational schooling: A systematic review*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation and Policy Development.
- Marsh, H. W. (1989). Effects of attending single-sex and coeducational high-schools on achievement, attitudes, behavior, and sex differences. *Journal of Educational Psychology*, 81, 70-85. doi:10.1037/0022-0663.81.1.70
- Martin, C. L., & Fabes, R. A. (2001). The stability and consequences of same-sex peer interactions. *Developmental Psychology*, 37, 431-446. doi:10.1037/0012-1649.37.3.431
- Melhuish, E. C., Sylva, K., Sammons, P., Siraj-Blatchford, I., Taggart, B., Phan, M. B., & Malin, A. (2008, August 29). Preschool influences on mathematics achievement. *Science*, 321, 1161-1162. doi:10.1126/science.1158808
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Boston, MA: Boston College, Lynch School of Education International Study Center.
- Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, . . . O'Connor, K. M. (2003). *TIMSS Assessment frameworks and specifications 2003* (2nd ed.). Boston, MA: Boston College, Lynch School of Education International Study Center.
- No Child Left Behind Act, 20 U.S.C. § 6301 et seq (2001).
- Novotney, A. (2011). Coed versus single-sex ed: Does separating boys and girls improve their education? Experts on both sides of the issue weigh in. *Monitor on Psychology*, 42, 58.
- Organization for Economic Cooperation and Development. (2010). *PISA 2009 results*. Retrieved from www.oecd.org/edu/pisa/2009
- Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.
- Pahlke, E. E., Bigler, R. S., & Patterson, M. M. (2012). *Reasoning about single-sex schooling for girls among students, parents, and teachers*. Manuscript submitted for publication.
- Park, H., & Behrman, J. (2010). *Causal effects of single-sex schools on college attendance: Random assignment in Korean high schools* (Population Studies Center Working Paper Series 10-01). Philadelphia, PA: University of Pennsylvania, Population Studies Center.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). London, England: Sage.
- Rudman, L. A., & Glick, P. (2008). *The social psychology of gender: How power and intimacy shape gender relations*. New York, NY: Guilford Press.
- Rustad, N., & Woods, J. (2004). Statement on the legality of single-sex education. Washington, DC: American Association of University Women. Retrieved from http://www.aauw.org/advocacy/issue_advocacy/actionpages/upload/singlesex_comments.pdf
- Sadker, M., & Sadker, D. (1994). *Failing at fairness: How our schools cheat girls*. New York, NY: Simon and Schuster.
- Sax, L. (2005). *Why gender matters*. New York, NY: Doubleday.
- Sax, L. (2010). Sex differences in hearing: Implications for best practice in the classroom. *Advances in Gender and Education*, 2, 13-21.
- Shapka, J. D., & Keating, D. P. (2003). Effects of a girls-only curriculum during adolescence: Performance, persistence, and engagement in mathematics and science. *American Educational Research Journal*, 40, 929-960. doi:10.3102/00028312040004929
- U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Retrieved from http://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp

Received March 30, 2012

Revision received October 5, 2012

Accepted November 26, 2012 ■

The Transition From Informal to Formal Mathematical Knowledge: Mediation by Numeral Knowledge

David J. Purpura
Purdue University

Arthur J. Baroody
University of Illinois at Urbana–Champaign

Christopher J. Lonigan
Florida State University

The purpose of the present study was to determine if numeral knowledge—the ability to identify Arabic numerals and connect Arabic numerals to their respective quantities—mediates the relation between informal and formal mathematical knowledge. A total of 206 3- to 5-year-old preschool children were assessed on 6 informal mathematics tasks and 2 numeral knowledge tasks. A year later, these children were assessed on 2 measures of formal mathematical knowledge, namely, the Woodcock-Johnson III Calculation Subtest and a formal number combinations task. Mediation analyses revealed that the relation between informal and formal mathematical knowledge is fully mediated by numeral knowledge, but only when both the skill of numeral identification and an understanding of numeral to quantity relations are considered.

Keywords: mathematics, preschool, learning trajectory, informal, formal

Mathematical knowledge is a critical aspect of early development (Baroody, Lai, & Mix, 2006; Jordan, Hanich, & Uberti, 2003). It provides a foundation for other academic abilities, as indicated by the strong predictive relation between early mathematics achievement and a broad range of later academic abilities (Duncan et al., 2007; Geary, 1994; Jordan, Kaplan, Ramineni, & Locuniak, 2009; National Mathematics Advisory Panel, 2008). Unfortunately, children who fall behind their peers early in mathematics usually continue to develop at a slower rate than more advanced peers and are likely to remain behind them (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004). The cumulative nature of early mathematics development—later competencies building on earlier ones—underscores the need to identify key early concepts and skills, determine how they develop, and elucidate their developmental relations. By understanding how these key mathematical skills and concepts are developmentally interrelated, effective classroom curricula and interventions can be constructed that might aid in reducing, or even eliminating, later mathematics difficulties.

Early Mathematics Learning Trajectories

The meaningful development of mathematical knowledge stems from constructing a systematic and well-interconnected web of mathematical concepts and skills (Baroody, 2003; Gersten & Chard, 1999; National Mathematics Advisory Panel, 2008). By connecting new information to previously learned knowledge, children are able to develop deep and flexible mathematical understanding (Hatano, 2003; James, 1958; Piaget, 1964). This often entails learning mathematical concepts and skills in an empirically delineated sequence. Such a sequence of the development of mathematical concepts and skills is called a learning trajectory (Clements, 2007; Clements & Sarama, 2004; Gravemeijer, 2002). Although significant work has been conducted to assess the developmental sequence of individual skills or concepts (Fuson, 1988; Gelman & Gallistel, 1978) and to construct learning trajectories of early mathematical knowledge (Clements & Sarama, 2009; Sarama & Clements, 2009), prior work has been primarily focused on the development within specific constructs (e.g., how children's counting ability develops across time) or on relating individual early skills to later skills (Aunio & Niemivirta, 2010; Clarke & Shinn, 2004; Martinez, Missall, Graney, Aricak, & Clarke, 2008; Muldoon, Lewis, & Freeman, 2009). What is needed is an empirically validated structure of how mathematics skills or concepts are related to broader transitions or phases of mathematical development.

A Key Transition in Early Mathematical Knowledge

The transition from informal everyday mathematical knowledge to formal school-taught mathematical knowledge is a particularly important juncture in mathematical development (Ginsburg, 1975; Greenes, Ginsburg, & Balfanz, 2004; Starkey, Klein, & Wakeley, 2004). This transition begins in preschool and kindergarten as children begin to learn the written numbering system and leads to

This article was published Online First March 18, 2013.

David J. Purpura, Department of Human Development and Family Studies, Purdue University; Arthur J. Baroody, College of Education, University of Illinois at Urbana–Champaign; Christopher J. Lonigan, Department of Psychology, Florida State University.

This work was supported by Institute of Education Science, U.S. Department of Education Grants R305B04074 to Christopher J. Lonigan and R305B100017 to Arthur J. Baroody. Views expressed herein are solely those of the authors and have not been reviewed or cleared by the grantors.

Correspondence concerning this article should be addressed to David J. Purpura, Department of Human Development and Family Studies, Purdue University, 1202 West State Street, Room 231, West Lafayette, IN 47907-2055. E-mail: purpura@purdue.edu

the eventual use of mathematics as a vehicle for acquiring and expressing knowledge in other domains such as science and engineering (Basista & Matthews, 2002). Discussed in turn are definitions of informal and formal mathematics, why this transition is critically important, how numeral knowledge develops, the role such development plays in formal mathematical development, and the rationale for the present study.

Informal Mathematics

Informal mathematical knowledge is composed of those competencies generally learned before or outside of school, often in spontaneous but meaningful everyday situations including play, and is characterized by the use of nonconventional and even self-invented symbols, strategies, or procedures rather than conventional written symbols or algorithms (Ginsburg, 1977). Although aspects of informal mathematical knowledge often do not require specific school-based instruction, these skills are malleable and can be enhanced through appropriate and targeted instruction (Arnold, Fisher, Doctoroff, & Dobbs, 2002; Baroody, Eiland, & Thompson, 2009; Clements & Sarama, 2007; Frye et al., 2013; Siegler & Ramani, 2008, 2009).

The central aspects of informal mathematics in the domain of number and operations are flexibly connecting quantities to number words and understanding the relations among these quantities. It has been hypothesized that children go through three overlapping levels of informal mathematics development (Krajewski, 2008; Krajewski & Schneider, 2009): Level 1 = *foundational skills*—distinguishing among quantities (comparison of sets of quantities) and learning to recite the verbal number word sequence (verbal counting) are foundational skills that develop separately; Level 2 = *meaningful numbering skills*—applying the count sequence to fixed sets via one-to-one counting and linking specific number words and quantities via cardinal number knowledge and verbal subitizing (understanding and representing that each number word represents a distinct quantity; e.g., “three” indicates ●●●); and Level 3 = *operations on verbal numbers*—understanding how actions on verbally represented numbers affect them (e.g., recognizing that the outcome of an addition word problem—not involving 0—is larger than either addend). The informal concepts and skills from each of these levels are central to many state and national standards for early mathematics (Common Core State Standards, 2011; National Council of Teachers of Mathematics, 2000, 2006), because they are the developmental precursors to understanding and learning formal mathematics (Bryant, Bryant, Kim, & Gersten, 2006; Chard et al., 2005; Cross, Woods, & Schweingruber, 2009; Frye et al., 2013; Geary, 1994; Ginsburg, Klein, & Starkey, 1998; Griffin & Case, 1997; Jordan et al., 2009; National Mathematics Advisory Panel, 2008).

Formal Mathematics

Formal mathematical knowledge consists of those skills and concepts taught in school and include the use of conventional written numerical notation (e.g., Arabic numerals and operation/equality signs) and written algorithms (e.g., multidigit addition with renaming; Ginsburg, 1977). One of the first—and a particularly important—formal school mathematics skill is fluency with verbally or graphically presented basic number com-

binations (i.e., addition and subtraction problems presented as a verbal or written expression or equation, such as “Two plus three,” “Three and two equals what?” and $2 + 3$ or $3 + 2 = \square$). The Common Core State Standards (2011)—and the curricula that have been aligned with these standards—specify that kindergarteners should be fluent with formal addition and subtraction combinations up to five. Fluency with these formal combinations is one of the first aspects of formal knowledge that children develop—typically beginning at the end of preschool and into kindergarten (National Mathematics Advisory Panel, 2008). The National Mathematics Advisory Panel (2008) indicated that children’s fluency with basic formal number combinations is critical to long-term mathematics achievement. Kindergartners’ fluency with formal number combinations is predictive of later mathematics achievement and learning difficulties (Jordan et al., 2009; Jordan & Levine, 2009; Mazzocco & Thompson, 2005). Children’s formal mathematics skills have also been shown to be malleable through targeted interventions (Baroody, Eiland, Purpura, & Reid, 2012, in press; Clarke et al., 2011; Fuchs et al., 2009). However, the mechanism by which the transition from informal to formal knowledge occurs is unclear. To design and implement targeted interventions for formal mathematical knowledge, it is necessary to know, not only what the key developmental precursors are, but how the developmental precursors connect in their development to formal mathematical knowledge.

The Relation Between Informal and Formal Mathematics

A sizeable body of research has found that different aspects of informal mathematics are important precursors to the acquisition of formal mathematics. For example, longitudinal work by Aunola et al. (2004) found that counting skills in preschool were predictive of both performance level and growth of mathematics skills through second grade. Similarly, research by Aunio and Niemivirta (2010) found that both early counting skills and quantity comparison skills in kindergarten were predictive of first grade general mathematics. Stock, Desoete, and Royers (2007) found that one-to-one counting was a significant marker of later mathematics difficulties, and VanDerHeyden, Broussard, and Cooley (2006) identified one-to-one counting in preschool as a strong predictor of kindergarten mathematics skills. Other work has found that the connection between number words and specific quantities (e.g., subitizing and cardinal number knowledge) is necessary for the attainment of more advanced mathematics (Kroesbergen, Van Luit, Van Lieshout, Loosbroek, & Van de Rijt, 2009; Palmer & Baroody, 2011; Sarnecka & Carey, 2008). Finally, Jordan, Kaplan, Locuniak, and Ramineni (2007) found that, among other variables, verbal story problems and number combinations at kindergarten accounted for a large amount of variance in predicting children’s mathematics skills as late as second grade. One concern with existing research is the assumption that informal skills have a direct effect on the development of formal mathematics skills. However, it is likely that another step in development, connecting informal knowledge to the written symbols, is necessary for the acquisition of formal knowledge because it provides a bridge between informal number and arithmetic knowledge and formal representations and procedures.

The Development of Numeral Knowledge

Although the acquisition of numeral names is a central part of the development of numeral knowledge—and presumably one of the more common “mathematical” activities done in school and home—it is critically important to recognize that the development of numeral knowledge extends beyond simply identifying or naming Arabic numerals. This development also involves connecting the written symbols to distinct quantities (e.g., the numeral “4” is a way to represent the quantity ●●●●). Children’s ability to identify written numerals and to connect written numerals with number words and quantities have been found to be strong, if not the strongest, predictors of later formal mathematics ability (Bryant et al., 2006; Clarke & Shinn, 2004; Chard et al., 2005; Griffin, Case, & Siegler, 1994; Lembke & Foegen, 2006; Lembke & Foegen, 2009). However, Baroody and Wilkins (1999) indicated that even though one of the first steps toward the development of formal knowledge is learning how to read and write numerals, numeral knowledge could not be classified as either formal or informal knowledge because it does not strictly meet the definition of either.

The development of aspects of numeral knowledge typically begins shortly after children start to develop aspects of their informal mathematical abilities such as knowledge of the counting sequence and the mapping of quantities onto number words (Krajewski & Schneider, 2009; Sarama & Clements, 2009). Once children understand and recognize that numerals are distinct representations from other symbols (e.g., numerals and letters are different), they are able to begin to connect numeral names with the written symbols. Approximately one-quarter of children can identify the numerals 1 to 9 by the time they turn 4 years old (Ginsburg & Baroody, 2003), and some children even begin to identify the first numerals (e.g., 1 and 2) when they are as young as 18 months (Fuson, 1988; Mix, 2009; Sarama & Clements, 2009).

Theoretical Role of Numeral Knowledge in Formal Mathematics Development

Both informal knowledge (connecting number words and quantities) and numeral knowledge (connecting number words and quantities to written symbols) have been shown to be independently related to formal mathematics knowledge (Aunola et al., 2004; Bryant et al., 2006; Clarke & Shinn, 2004). However, some research has indicated that numeral knowledge acts as a ceiling on general mathematical knowledge, preventing children from completing mathematical tasks above their level of numeral knowledge (Sinclair, Siegrist, & Sinclair, 1983). It also has been found that many children with mathematics disabilities tend to have specific difficulties with the symbolic numeral system (Rousselle & Noel, 2007), rather than with informal knowledge (Song & Ginsburg, 1987) or with other cognitive domains (Butterworth & Reigosa, 2007)—indicating that a deficit in an aspect of numeral knowledge development (or deficits in both aspects of numeral knowledge) may inhibit children’s successful acquisition of formal mathematics. As such, numeral knowledge may act as a mediator in the development from informal to formal mathematical knowledge.

Rationale for the Present Study

Both informal knowledge and numeral knowledge have been found to be strong predictors of later mathematics achievement and appear to play key roles in the development of formal mathematics. Baroody and Ginsburg (1990) suggested that children who fail to successfully make this transition are at significant risk for later learning difficulties, even if their informal mathematical knowledge had been developing typically. However, the means by which children’s informal knowledge and numeral knowledge contribute to the development of formal mathematics is not entirely clear. Thus, the goal of the present study was to determine if informal knowledge directly contributes to the development of formal knowledge or if this relation is mediated in some fashion by numeral knowledge. Further, a secondary goal was to determine the nature of such mediation by testing to see whether (a) the mapping of number word names (which themselves have already been connected to quantities) onto the symbols alone was sufficient to make the connection between informal and formal mathematics, (b) the connection between the quantities (which have already connected with the number words) and the symbols is sufficient to make the connection between informal and formal skills, or (c) both are necessary to make the connection between informal and formal mathematics. In aligning with the theoretical development presented earlier, it was hypothesized that numeral knowledge would fully mediate the relation between informal and formal mathematics and that the combination of both of the numeral knowledge components would be necessary to achieve full mediation.

Method

Participants

In the first year of this study, data were collected from 393 preschool children in 44 public and private preschools serving children from families of low to middle socioeconomic status living in Northern Florida. In the second year of the study, 206 of the original children were tested again. Of these children, 112 had moved on to kindergarten and attended 28 different public or private elementary schools in two counties. The other 94 children moved on to their second year in preschool at 20 different public (Head Start) or private preschools in two counties. The 206 children who completed the assessments at both time points were evenly split by sex (51.9% female) and approximately representative of the demographics of Northern Florida (60.2% Caucasian, 28.2% African American, and 11.6% other race/ethnicity). At Time 1, children ranged in age from 3.18 years to 5.88 years ($M = 4.66$ years, $SD = 0.69$ years). The children were primarily English speaking and had no known developmental disorders. The children who completed both testing points were not significantly different on any of the Time 1 mathematics variables than the children who did not complete the Time 2 assessment. Parental consent was obtained for each participating child.

Materials

Children were assessed on informal, numeral, and formal knowledge tasks. These tasks were assessed as a part of a larger

battery of tests that took approximately three 20- to 30-min sessions at each time point. All tasks (except an achievement test that was one of the two tasks used to assess formal knowledge) were developed as part of a broader measure (Purpura, 2010). Items on each of the tasks from the broader measure were derived by a process using item response theory that ensured that each item was related to its intended construct (e.g., set comparison), had adequate discrimination (*a* parameter), and did not duplicate the difficulty level (*b* parameter) of other items on the same task. Raw total scores were used for each measure.

Tasks Administered at Time 1

Informal mathematics tasks. Six tasks served to measure informal knowledge.

Verbal counting. Children were asked to count as high as possible. When a child made a mistake, or correctly counted to 100 without making a mistake, the task was stopped. Spontaneous self-corrections were not scored as incorrect, and the child was allowed to continue counting. The highest number counted to was converted to a score based on a 7-point scale. Children were awarded one point each for correctly counting to 5, 10, 15, 20, 25, 40, and 100.

One-to-one counting. Children were presented with a set of three, six, 11, 14, or 16 dots on a page and asked to count the set. Children were awarded one point for each set if they correctly counted each dot only once. This task had an internal consistency (Chronbach's alpha) of .79.

Cardinality. This task was assessed in the context of the one-to-one counting task. At the completion of the counting three, six, and 11 one-to-one counting items, children were asked to indicate how many dots there were in all. Children were awarded one point if they restated the last number counted ("how many?"). This task had an internal consistency of .75.

Subitizing. Children were briefly presented (2 s) a set of pictures (set sizes from one to seven presented in a linear fashion; e.g., ●●●●) and instructed to say how many dots or pictures were presented. For each correct response, children were awarded one point. This task had an internal consistency of .69.

Set comparison. For each of the six items, children were presented with four sets of dots on a page representing different quantities (e.g., | ●●● | ●● | ●●●● | ● |). They were then asked which set had the most (three items) or fewest dots (three items). Children received one point for pointing to the correct set. This task had an internal consistency of .77.

Story problems. Children were presented verbally with story problems that did not contain distracters (e.g., irrelevant information). These story problems were simple addition (three items) or subtraction problems (four items) that were appealing to children. For example, one question was, "Johnny had one cookie and his mother gave him one more cookie, how many cookies did he have now?" Children were awarded one point for each correct response. This task had an internal consistency of .71.

Numeral knowledge skill tasks. Two tasks served to measure numeral knowledge.

Numeral identification. Children were presented with flashcards of nine numbers (1, 2, 3, 7, 8, 10, 12, 14, and 18). They were shown the flashcards one at a time and asked, "What number is this?" For each correct response children were awarded one point. This task had an internal consistency of .90.

Set to numerals. On the first three items in this task, children were presented with a numeral at the top of the page (e.g., 3) and five sets of dots below (e.g., | ●●●● | ● | ●●● | ●● | ●●●●● |). They were instructed to identify which of the sets meant the same thing as the number at the top of the page. On the last two items of this task, children were presented with a set of dots at the top of the page (e.g., ●●●●) and five numerals at the bottom (e.g., 4, 2, 3, 1, 5). They were instructed to identify which of the numerals meant the same thing as the set of dots at the top of the page. Children were awarded one point for each correct response. This task had an internal consistency of .80.

Tasks Administered at Time 2

Formal mathematics skills tasks. Two tasks were used to gauge basic formal knowledge. These tasks were selected because they could be used to assess the most basic addition and subtraction combinations. The primary differences between these tasks are in the presentation of the items and the method of response. In the first task, children are both shown and told the problem, for which they give a verbal answer. In the second task, children are just shown the problem, and they are asked to give a written response.

Number combinations. Children were presented with a formal addition problem (e.g., $1 + 1 =$) and asked, "How much is . . . [stated the problem]." There were five total problems: $0 + 2 =$, $1 + 1 =$, $1 + 2 =$, $2 + 2 =$, $1 + 3 =$. For each correct response children were awarded one point. This task had an internal consistency of .77. This measure was also administered in the first year of the study; however, overall performance on this task was low ($M = 1.19$, $SD = 1.50$) suggesting that the majority of children had little to no formal knowledge at Time 1.

Woodcock-Johnson III Calculation subtest (WJ-III Calc). The WJ-III Calc subtest is a paper-and-pencil arithmetic test where children are asked to solve addition and subtraction problems and has been shown to have a median reliability of .92 for children 5–19 years old (Woodcock, McGrew, & Mather, 2001). Children were awarded one point for each correct answer.

Procedure

Assessment procedure. Preschoolers were assessed on the informal mathematics tasks and numeral knowledge tasks in the spring of Year 1. Participants were assessed a year later (spring of Year 2) on their formal knowledge when slightly over half of the children had advanced to kindergarten. Individuals who had either completed or were working toward completion of a bachelor's degree conducted the assessments. The assessors each completed a 2- to 3-hr training on the measures prior to each testing point (Time 1 and Time 2) and completed an extensive testing out process to ensure accuracy of administration. Assessments took place in the local preschools or kindergarten classrooms during noninstructional time in a quiet room designated by the individual school directors or teachers.

Analytic procedure. As the primary analytic method was to conduct mediation analyses, data analysis was conducted in five steps based on the recommendations of Baron and Kenny (1986) in conjunction with updated recommendations by Zhao, Lynch, and Chen (2010). The first two steps were analyses of the direct effects of informal mathematical knowledge on (Step 1) numeral knowl-

edge and (Step 2) formal mathematical knowledge. The third step was an analysis of the direct effects of numeral knowledge on formal knowledge when controlling for informal knowledge. The fourth step was an evaluation of the mediation effects of numeral knowledge on the relation between informal and formal mathematical knowledge using the percentile bootstrap approach (recommended by Zhao et al., 2010) rather than the Sobel test (recommended by Baron & Kenny, 1986), because the percentile bootstrap approach is more powerful in detecting mediation effects than the Sobel test (Preacher & Hayes, 2004). The fifth step was a comparison between the baseline model from step four and a simpler model that did not include the direct effects of informal mathematical knowledge on formal mathematical knowledge. To evaluate whether only one of the numeral knowledge variables by itself was sufficient to mediate the relation between informal and formal knowledge, these same steps were repeated two additional times, replacing the numeral knowledge latent factor with the individual numeral knowledge variables. All models in the analyses were logically identified.

Results

Descriptive Statistics

Means, standard deviations, skewness, and kurtosis for all variables are included in Table 1 and are presented by age group (younger children are those children who were still in preschool at Time 2, and older children are those children who were in kindergarten at Time 2) in Table 2. All data are presented as raw scores. The distributions of scores for all variables in this study were normal.¹ Correlations between the mathematics tasks that were assessed are presented in Table 3. No significant gender differences in preschool mathematics scores were found. When analyses were conducted using age-regressed standard scores, the results were comparable to the analyses conducted with raw scores. As such, the results using raw scores are reported because the scores are more interpretable.

Primary Analysis

Step 1: Direct effects of informal mathematical knowledge on numeral knowledge. Analyses of the relation between informal mathematical knowledge and numeral knowledge indicated that informal mathematical knowledge significantly predicted numeral knowledge ($\beta = 0.94, p < .001$).

Step 2: Direct effects of informal mathematical knowledge on formal mathematical knowledge. Analyses of the relation between informal mathematical knowledge and formal mathematical knowledge indicated that informal mathematical knowledge significantly predicted formal mathematical knowledge ($\beta = .84, p < .001$).

Step 3: Direct effects of numeral knowledge on formal mathematical knowledge. Analysis of the relation between numeral knowledge and formal mathematical knowledge, when controlling for the effects of informal mathematical knowledge, indicated that numeral knowledge significantly predicted formal mathematical knowledge ($\beta = .86, p = .038$).

Step 4: Mediation effects of numeral knowledge on the relation between informal and formal mathematical knowledge. Significant mediation effects were determined through use of the percentile bootstrap approach. This method

Table 1
Means, Standard Deviations, Range, Skewness, and Kurtosis of the Sum Scores of the Mathematics Tasks

Task	<i>M</i>	<i>SD</i>	Range ^a	Skew	Kurtosis
Testing Time 1					
Informal mathematics					
Verbal counting	3.43	1.83	0–7	.27	–1.01
One-to-one counting	3.33	1.50	0–5	–.41	–0.96
Cardinality	2.18	1.02	0–3	–.96	–0.34
Subitizing	3.86	1.61	0–7	–.26	0.01
Set comparison	3.92	1.89	0–6	–.44	–1.04
Story problems	3.36	1.96	0–7	.11	–0.90
Numeral knowledge					
Number identification	5.35	3.06	0–9	–.36	–1.17
Set to numerals	2.94	1.67	0–5	–.36	–1.10
Testing Time 2					
Formal mathematics					
WJ-III calculation	3.62	3.72	0–16	.87	0.15
Number combinations	2.87	1.96	0–5	–.27	–1.51

Note. *N* = 206. WJ-III = Woodcock-Johnson Tests of Achievement (3rd ed., Woodcock, McGrew, & Mather, 2001).
^a The range indicates both the possible and actual range of scores for all tasks other than the WJ-III Calculation task. For the latter task, only the actual range is presented.

utilizes a random sampling with replacement approach to calculate a sampling of indirect effects. Indirect effects from each sampling were then sorted from low to high and the highest and lowest 2.5% (when using a 95% confidence interval) were removed. Significant mediation effects are present if the confidence interval does not contain 0. Mediation analyses showed significant mediation effects of numeral knowledge on the relation between informal and formal mathematical knowledge. In Figure 1, the mediation model that includes a direct—but nonsignificant—effect of informal mathematical knowledge on formal mathematical knowledge is presented. Overall, the magnitude of the indirect effect of the mediation model (or amount of variance accounted for in formal mathematical knowledge by the indirect effect) was large ($R^2 = .81$) and accounted for 98% of the total variance.

Step 5: Comparison of models to determine full or partial mediation. The model with the direct effect of informal mathematical knowledge on formal mathematical knowledge was then compared to the same model without the direct effect of informal mathematical knowledge on formal mathematical knowledge included. In Table 4, the tests of model fits for both models are presented. Test of chi-square differences indicated no significant difference between the models and all other fit indices were nearly identical between the models. Given that no differences were found between the model fits, the more parsimonious model (the full mediation model—the one with no direct effect of informal knowledge on formal knowledge) was selected as the preferred model. In general, the model fit indices provide evidence that the selected model provides a good fit to the data (Brown, 2006; Hu & Bentler, 1999; Mueller & Hancock, 2010).

¹ All children were able to attempt the items on the WJ-III Calculation test. Of the 206 children who participated in the study, 72 obtained a raw score of 0 on the WJ-III Calculation test and 40 obtained a score of 0 on the number combinations task. Importantly, only 27 total children (13%) obtained scores of zero on *both* tasks.

Table 2
Means, Standard Deviations, and Ranges of the Sum Scores of the Mathematics Tasks by Age Group

Task	Younger children			Older children		
	<i>M</i>	<i>SD</i>	Range ^a	<i>M</i>	<i>SD</i>	Range ^a
Informal mathematics						
Verbal counting	2.49	1.35	0–7	4.20	1.81	0–7
One-to-one counting	2.81	1.49	0–5	3.75	1.37	0–5
Cardinality	1.78	1.11	0–3	2.51	0.80	0–3
Subitizing	3.33	1.58	0–7	4.30	1.52	0–7
Set comparison	3.12	1.78	0–6	4.58	1.73	0–6
Story problems	2.52	1.66	0–6	4.05	1.93	0–7
Numeral knowledge						
Number identification	4.06	3.01	0–9	6.41	2.68	0–9
Set to numerals	2.05	1.56	0–5	3.66	1.39	0–5
Formal mathematics						
WJ-III calculation	1.13	2.08	0–10	5.67	3.53	0–16
Number combinations	1.65	1.60	0–5	3.88	1.64	0–5

Note. *N* = 206. WJ-III = Woodcock-Johnson Tests of Achievement (3rd ed., Woodcock, McGrew, & Mather, 2001). Younger children *n* = 93, older children *n* = 113.

^a The range indicates both the possible and actual range of scores for all tasks except the story problems task for the younger children and the WJ-III calculation task for both the younger and older children. For the story problems task, the maximum possible correct was seven. The WJ-III calculation subtest is design for individuals of all ages and thus, the scores presented only represent the actual range attained in this sample.

Mediation by Individual Numeral Knowledge Variables

The same series of mediation analyses were conducted using each of the numeral knowledge variables separately to determine if one aspect of numeral knowledge accounted for the mediation findings.

Step 1: Direct effects of informal mathematical knowledge on numeral knowledge variables. Analyses of the relation between informal mathematical knowledge and performance on the numeral identification task and performance on the set-to-

numerals task indicated that informal mathematical knowledge significantly predicted performance on the numeral identification task ($\beta = .74, p < .001$) and significantly predicted performance on the set-to-numeral task ($\beta = .76, p < .001$).

Step 2: Direct effects of informal mathematical knowledge on formal mathematical knowledge. The results in this step are the same as were reported in the previous Step 2. Analyses of the relation between informal mathematical knowledge and formal mathematical knowledge indicated that knowledge of informal

Table 3
Correlations Between the Sum Scores of All the Mathematical Knowledge Tasks

Variable	1	2	3	4	5	6	7	8	9	10
Testing Time 1										
Informal mathematics										
1. Verbal counting	—									
2. One-to-one counting	.62	—								
3. Cardinality	.55	.72	—							
4. Subitizing	.41	.49	.43	—						
5. Set comparison	.52	.52	.51	.35	—					
6. Story problems	.52	.46	.49	.46	.62	—				
Numeral knowledge										
7. Numeral identification	.55	.54	.55	.45	.55	.57	—			
8. Set to numerals	.60	.59	.54	.48	.58	.53	.65	—		
Testing Time 2										
Formal mathematics										
9. WJ-III calculation	.53	.46	.47	.39	.52	.60	.54	.52	—	
10. Number combinations	.53	.49	.49	.42	.54	.55	.62	.63	.68	—

Note. *N* = 206. WJ-III = Woodcock-Johnson Tests of Achievement (3rd ed., Woodcock, McGrew, & Mather, 2001). All correlations were significant at $p < .01$.

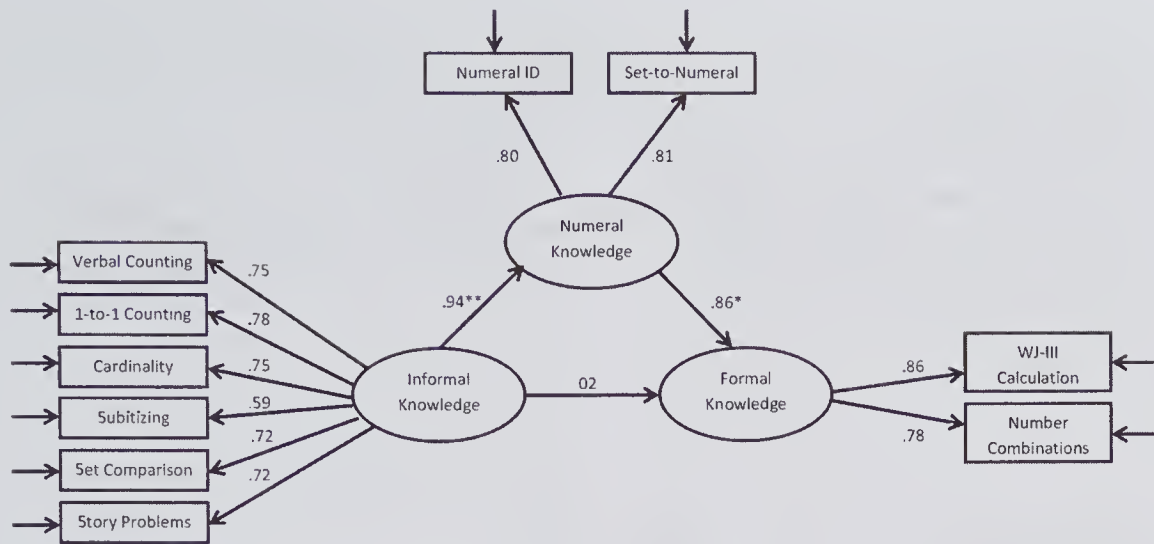


Figure 1. The figure shows the mediation of numeral knowledge in the relation between informal and formal mathematical knowledge. WJ-III = Woodcock-Johnson Tests of Achievement (3rd ed.; Woodcock, McGrew, & Mather, 2001). * $p < .05$. ** $p < .01$.

mathematical knowledge significantly predicted formal mathematical knowledge ($\beta = .84, p < .001$).

Step 3: Direct effects of individual numeral knowledge variables on formal mathematical knowledge. Analysis of the relation between performance on the numeral identification task and formal mathematical knowledge and between performance on the set-to-numerals task and formal mathematics, when controlling for the effects of informal mathematical knowledge, indicated that performance on the numeral identification task significantly predicted formal mathematical knowledge ($\beta = .20, p = .029$) and performance on the set-to-numerals task marginally significantly predicted formal mathematical knowledge ($\beta = .17, p = .079$).

Step 4: Mediation effects of individual numeral knowledge variables on the relation between informal and formal mathematical knowledge. Significant mediation effects were determined through use of the percentile bootstrap approach. Mediation analyses showed significant mediation effects of performance on the numeral identification task on the relation between informal and formal mathematical. Overall, the magnitude of the indirect effect of the mediation model was ($R^2 = .15$) and accounted for only 21% of the total variance. Mediation analyses also showed a marginally significant mediation effect of performance on the set-to-numerals task. Overall, the magnitude of the indirect effect of the mediation model was small ($R^2 = .13$) and accounted for only 19% of the total variance. These findings revealed that

individually, both performance on the numeral identification task and performance on the set-to-numerals task only partially mediated the relation between informal and formal mathematical knowledge because the direct effect of informal knowledge on formal knowledge remained large and significant (see Figures 2A and 2B). In fact, in both of these analyses, informal knowledge accounted for the majority of the variance in formal knowledge. These findings suggest that the relation between informal and formal mathematical knowledge is fully mediated by numeral knowledge but only when both aspects of numeral knowledge are considered together.

Discussion

The results of this study indicate that the relation between informal and formal mathematical knowledge is fully mediated by children's numeral knowledge. Mapping both number-words and quantities to the written symbols are necessary steps for children to apply their formal mathematics knowledge to formal concepts. Although prior research typically tied informal knowledge directly to the development of formal knowledge (Aunola et al., 2004; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Stock, Desoete, & Royers, 2007), the current findings indicate that there is no direct impact of informal mathematical knowledge on formal mathematical knowledge. Rather, children must map their informal knowledge directly onto numeral knowledge, which then must be

Table 4
Fit Indices for the Mediation Models With and Without the Direct Effect of Informal Mathematical Knowledge on Formal Mathematical Knowledge

Model	χ^2	df	CFI	TLI	RMSEA	SRMR	χ^2 dif
With direct effect	84.29	32	.95	.94	.09	.04	—
Without direct effect	84.29	33	.96	.94	.09	.04	0.00 (ns)

Note. $N = 206$. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual. The dash indicates that, because the comparison was made between the two models (e.g., the model without the direct effect was compared to the model with the direct effect), there was only one analytic comparison.

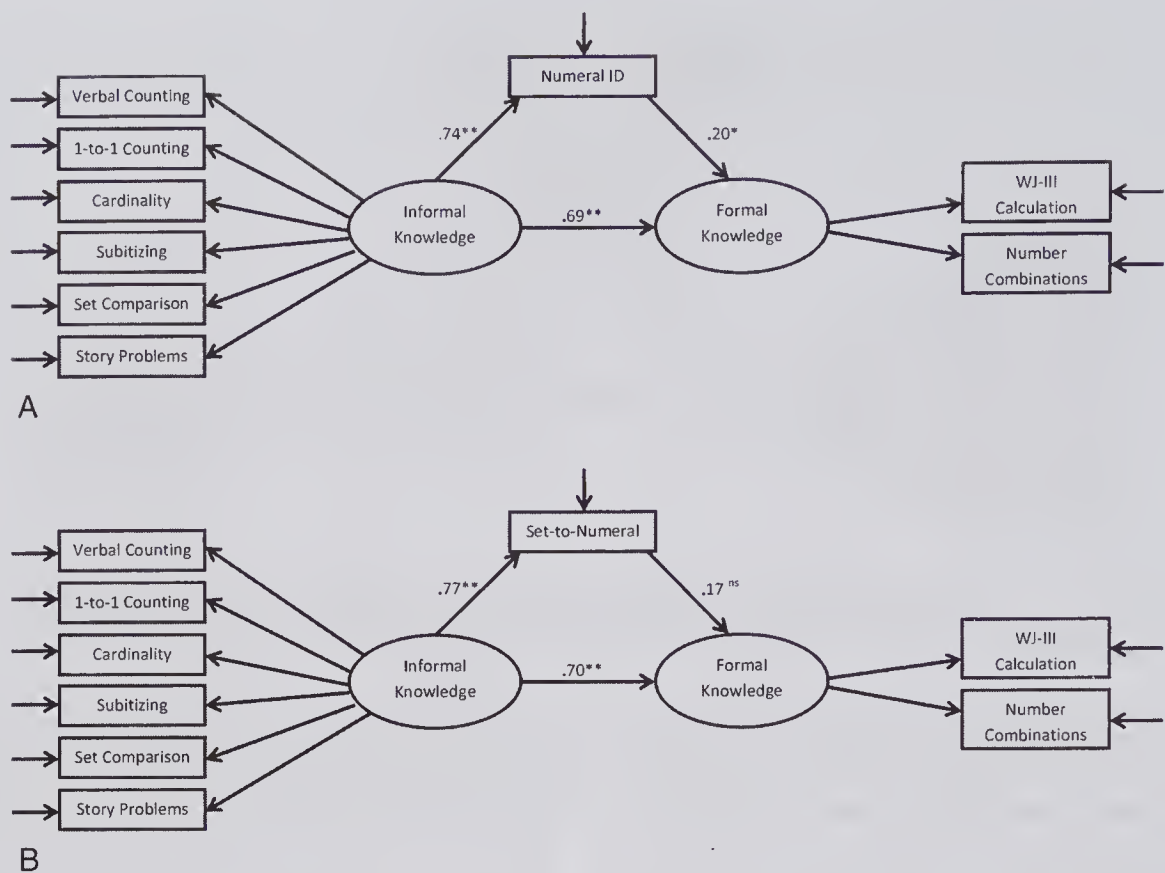


Figure 2. A. The partial mediation of numeral identification on the relation between informal mathematical knowledge and formal mathematical knowledge. B. The partial mediation of set-to-numerals task on the relation between informal mathematical knowledge and formal mathematical knowledge. WJ-III = Woodcock-Johnson Tests of Achievement (3rd ed.; Woodcock, McGrew, & Mather, 2001). * $p < .05$. ** $p < .01$.

mapped onto formal knowledge—suggesting that there is a developmental trajectory of these skills/concepts. Further, these results indicate that a depth of numeral knowledge—both the ability to identify numerals and connect numerals to quantities—is necessary to achieve the full mediation. Only partial mediation was found when either the numeral identification task or the set-to-numerals task was included in the model alone. This suggests that both the procedural skill of identifying numerals and the conceptual ability of understanding that each numeral represents a specific quantity appear to represent necessary functions of the numeral knowledge domain—meaning, numeral knowledge may act as a gatekeeper (or barrier) in the development of formal mathematical knowledge.

These findings delineate an important learning trajectory underscoring children’s mathematical growth across a critical developmental juncture by building on prior models of informal mathematics development (Krajewski & Schneider, 2009). Notably, these findings support a critical additional step not included in prior models of early mathematics acquisition. The mediational findings also help to explain why numeral knowledge skills are often found to be even more highly correlated with formal mathematics skills than are informal mathematics skills (with $r_s > .60$; Bryant et al., 2006; Clarke & Shinn, 2004; Lembke & Foegen, 2006, 2009; Purpura & Lonigan, 2013). As children must map informal knowledge onto numeral knowledge and then onto formal knowledge, it appears that numeral knowledge is essentially a necessary precursor for the acquisition of formal mathematical knowledge (Baroody & Wilkins, 1999). The results of this study

are not intended to discount the importance of the development of individual informal mathematics skills but, rather, put this development within a broader context because prior research has indicated that individual informal mathematics skills are directly connected to formal mathematical knowledge. For example, Aunola et al. (2004) utilized various measures of counting knowledge to predict later general mathematics knowledge and found large and significant relations between counting skills and general mathematics skills. Such significant effects would be expected given that the different counting skills are *steps* in the development of overall mathematical abilities. However, it is critical to emphasize that development of individual skills are *steps* in the broader acquisition of mathematics and the findings of this study clarify one critical step in the broader context.

Not only do these findings build on prior models of mathematics development and support Baroody and Wilkins’ (1999) assertion, the findings also fit the developmental framework and the theoretical structure indicated by theories of meaningful mathematical learning (Baroody, 1987; James, 1958; Piaget, 1964). In such views, it is believed that children must be able to connect each type of new mathematical information to existing knowledge to develop their mathematical competence. For preschool and kindergarten children to develop formal knowledge, they must not only learn the numerical symbols by which the system is structured (a procedural skill) but must actively connect the symbol name information to their informal knowledge of number-words and quantities (a conceptual ability). Essentially, children must develop the connection between number words and quantities, and then connect both the

number words and quantities to the written symbols to develop formal knowledge. This supports the growing recognition for the need to integrate procedural and conceptual knowledge (see Baroody, Feil, & Johnson, 2007; Cross et al., 2009; National Mathematics Advisory Panel, 2008). The results may also help to explain Mazzocco and Thompson (2005) findings that a composite of three informal items at kindergarten (cardinality, comparisons of one-digit numbers, and mentally adding one-digit numbers) and one numeral knowledge item (reading numerals) from the TEMA-2 were predictive of which children would be designated mathematically “learning disabled” during both Grades 2 and 3 because the composite included items that tapped informal, formal, and numeral knowledge abilities. Utilizing the developmental framework identified here, may lead to a better and more efficient process for early identification of children at risk of later mathematics difficulties.

Limitations

Two limitations of this study should be noted. First, there was significant attrition across the two time points, primarily due to student mobility. Although these data were assumed to be missing at random, and there were no differences on Time 1 mathematics scores between completers and noncompleters, the level of attrition could have added a level of unknown variation into the findings. Second, this study solely focused on the “exact language-based” number system and does not incorporate measures that assess the approximate (nonverbal) number system (ANS; Dehaene, 1992).

Future Directions

Identifying this developmental sequence of early mathematics skills provides a foundation on which to conduct future research. Specifically, there is a need to expand beyond the broad definitions of “informal” and “formal” mathematics and delineate how the individual informal or formal skills and concepts (e.g., one-to-one counting, comparison, or number combinations, place value) interact in their development to create the web of informal or formal mathematical knowledge. This learning trajectory can also be used to provide information for the development of targeted interventions. By understanding where in the learning trajectory a child’s skills are underdeveloped, teachers can provide specific interventions to enhance those skills and hopefully prevent future learning difficulties. Further, delineating such a sequence also would enable teachers to easily identify the next instructional phase for a typically performing or advanced student. A learning trajectory will also enable both teachers and researchers to identify whether children have developed the appropriate developmental prerequisites to benefit from a broader curriculum or specific intervention. For example, teachers may find that children who have not fully developed their understanding of numeral knowledge may not be ready for mathematics interventions involving formal knowledge. Conversely, they may find that a younger child who has developed both informal and numeral knowledge may be ready for more advanced instruction in formal concepts.

The expansion of this learning trajectory beyond the verbal/symbolic mathematics system could be important to developing a broader understanding of early mathematics development. Specif-

ically, identifying the role that the ANS plays—as it is related to informal and formal mathematics development—in early mathematics development may allow for better identification of early mathematics difficulties. The ANS may also play a direct or indirect role in contributing to the development of formal mathematics development (Gillmore, McCarthy, & Spelke, 2007; Libertus, Feigenson, & Halberda, 2011). Prior research has shown that the ANS is correlated with informal and formal mathematics skills, even after controlling for language and intelligence (Libertus et al., 2011); however, these relations were not evaluated controlling for other early mathematics abilities. Thus, it is not clear if the relation between the ANS and formal mathematics is a direct relation or an indirect relation mediated by informal knowledge. Future research should be conducted to evaluate the relation of the ANS to the current learning trajectory to better understand the broader development of children’s early mathematical concepts.

An additional direction for future research should be the determination of nonmathematical factors that account for the remaining variance in the developmental model. It is likely that additional variance could be accounted for by including cognitive or behavioral abilities in the model. Prior research has found that working memory (Swanson, 2004; Swanson & Beebe-Frankenberger, 2004; Swanson & Kim, 2007), attention (Fuchs et al., 2005, 2006), and rapid digit naming or processing speed (Cirino, 2011; Krajewski & Schneider, 2009) have an impact on the development of mathematical abilities. Further, children’s language and print knowledge skills have also been found to be important factors in mathematics development (Fuchs et al., 2008; Leong & Jerred, 2001; Purpura, Hume, Sims, & Lonigan, 2011). Given that both language and print knowledge have been identified as significant predictors in later reading development (Lonigan, Schatschneider, & Westberg, 2008; Morris et al., 1998; Stanovich, Siegal, & Gottardo, 1997)—and numeral identification is likely to be highly rooted in basic concepts of print knowledge and/or language development (LeFevre et al., 2010)—it is plausible that a child who does not adequately develop print and language skills will also not sufficiently develop their numeral knowledge. The precise stage in the learning trajectory model where cognitive, behavioral, and language/print skills affect development should be identified to determine if such connections play a role in combined mathematics and reading disorders. Ideally, it should be evaluated whether the impact is primarily found at one developmental level (e.g., informal mathematics), or whether the impact is general to all stages of mathematical development. If one of these specific factors is found to adversely impact the development of mathematics skills at either a specific or general level, it may be prudent to conduct interventions that target that factor (e.g., working memory, attention, language/print knowledge) in conjunction with early mathematics interventions to best improve children’s early mathematics skills.

References

- Arnold, D. H., Fisher, P. H., Doctoroff, G. L., & Dobbs, J. (2002). Accelerating math development in Head Start classrooms. *Journal of Educational Psychology, 94*, 762–770. doi:10.1037/0022-0663.94.4.762
- Aunio, P., & Niemivirta, M. (2010). Predicting children’s mathematical performance in grade one by early numeracy. *Learning and Individual Differences, 20*, 427–435. doi:10.1016/j.lindif.2010.06.003

- Aunola, K., Leskinen, E., Lerkkanen, M., & Nurmi, J. (2004). Developmental dynamics of math performances from preschool to Grade 2. *Journal of Educational Psychology*, 96, 699–713. doi:10.1037/0022-0663.96.4.699
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182. doi:10.1037/0022-3514.51.6.1173
- Baroody, A. J. (1987). *Children's mathematical thinking: A developmental framework for preschool, primary, and special education teachers*. New York, NY: Teachers College Press.
- Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 1–34). Mahwah, NJ: Erlbaum.
- Baroody, A. J., Eiland, M. D., Purpura, D. J., & Reid, E. E. (2012). Fostering at-risk kindergarten children's number sense. *Cognition and Instruction*, 30, 435–470. doi:10.1080/07370008.2012.720152
- Baroody, A. J., Eiland, M. D., Purpura, D. J., & Reid, E. E. (in press). Can discovery learning foster first graders' fluency with the most basic addition combinations? *American Educational Research Journal*.
- Baroody, A. J., Eiland, M., & Thompson, B. (2009). Fostering at-risk preschoolers' number sense. *Early Education and Development*, 20, 80–128. doi:10.1080/10409280802206619
- Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education*, 38, 115–131.
- Baroody, A. J., & Ginsburg, H. P. (1990). Children's mathematical learning: A cognitive view. *Journal for Research in Mathematics Education*, 21(Monograph No. 4), 79–90.
- Baroody, A. J., Lai, M., & Mix, K. S. (2006). Development of young children's early number and operation sense and its implications for early childhood education. In B. Spodek & O. N. Saracho (Eds.), *Handbook of research on the education of young children* (2nd ed., pp. 187–221). Mahwah, NJ: Erlbaum.
- Baroody, A. J., & Wilkins, J. L. M. (1999). The development of informal counting, number, and arithmetic skills and concepts. In J. V. Copley (Ed.), *Mathematics in the early years* (pp. 48–65). Washington, DC: National Association for the Education of Young Children.
- Basista, B., & Matthews, S. (2002). Integrated science and mathematics professional development programs. *School Science and Mathematics*, 102, 359–370. doi:10.1111/j.1949-8594.2002.tb18219.x
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Bryant, D. P., Bryant, B. R., Kim, S. A., & Gersten, R. (2006, February). *Three-tier mathematics intervention: Emerging model and preliminary findings*. Poster presented at the 14th annual meeting of the Pacific Coast Research Conference, San Diego, CA.
- Butterworth, B., & Reigosa, V. (2007). Information processing deficits in dyscalculia. In D. B. Berch & M. M. M. Mazzocco (Eds.), *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities* (pp. 65–81). Baltimore, MD: Brookes.
- Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, 30, 3–14. doi:10.1177/073724770503000202
- Cirino, P. T. (2011). The interrelationships of mathematical precursors in kindergarten. *Journal of Experimental Child Psychology*, 108, 713–733. doi:10.1016/j.jecp.2010.11.004
- Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33, 234–248.
- Clarke, B., Smolkowski, K., Baker, S. K., Fien, H., Doebler, C. T., & Chard, D. J. (2011). The impact of a comprehensive Tier 1 core kindergarten program on the achievement of students at risk in mathematics. *The Elementary School Journal*, 111, 561–584. doi:10.1086/659033
- Clements, D. H. (2007). Curriculum research: Toward a framework for "research-based curricula". *Journal for Research in Mathematics Education*, 38, 35–70.
- Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, 6, 81–89. doi:10.1207/s15327833mtl0602_1
- Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, 38, 136–163.
- Clements, D. H., & Sarama, J. (2009). *Learning and teaching early math: The learning trajectories approach*. New York, NY: Routledge.
- Common Core State Standards. (2011). *Common Core State Standards: Preparing America's students for college and career*. Retrieved from <http://www.corestandards.org/>
- Cross, C. T., Woods, T. A., & Schweingruber, H. (Eds.). (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Washington, DC: Committee on Early Childhood Mathematics, National Research Council.
- Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, 44, 1–42. doi:10.1016/0010-0277(92)90049-N
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, 43, 1428–1446. doi:10.1037/0012-1649.43.6.1428
- Frye, D., Baroody, A. J., Burchinal, M., Carver, S. M., Jordan, N. C., & McDowell, J. (2013). *Teaching math to young children: A practice guide*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Craddock, C. F., & Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology*, 100, 829–850. doi:10.1037/a0012657
- Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, 97, 493–513. doi:10.1037/0022-0663.97.3.493
- Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., . . . Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, 98, 29–43. doi:10.1037/0022-0663.98.1.29
- Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., . . . Zumeta, R. O. (2009). Remediating number combination and word problem deficits among students with mathematical difficulties: A randomized control trial. *Journal of Educational Psychology*, 101, 561–576. doi:10.1037/a0014701
- Fuson, K. C. (1988). *Children's counting and concepts of number*. New York, NY: Springer-Verlag. doi:10.1007/978-1-4612-3754-9
- Geary, D. C. (1994). *Children's mathematical development: Research and practical applications*. Washington, DC: American Psychological Association. doi:10.1037/10163-000
- Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Oxford, England: Harvard University Press.
- Gersten, R., & Chard, D. (1999). Number sense: Rethinking arithmetic instruction for students with mathematic difficulties. *The Journal of Special Education*, 33, 18–28. doi:10.1177/002246699903300102

- Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, 447, 589–591. doi:10.1038/nature05850
- Ginsburg, H. P. (1975). Young children's informal knowledge of mathematics. *Journal of Children's Mathematical Behavior*, 1, 63–156.
- Ginsburg, H. P. (1977). *Children's arithmetic: The learning process*. Oxford, England: Van Nostrand.
- Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability* (3rd ed.). Austin, TX: Pro-Ed.
- Ginsburg, H. P., Klein, A., & Starkey, P. (1998). The development of children's mathematical thinking: Connecting research with practice. In D. Williams, I. E. Sigel, & K. Renninger (Eds.), *Child psychology in practice* (pp. 401–476). Hoboken, NJ: Wiley.
- Gravemeijer, K. (2002, April). *Learning trajectories and local instruction theories as a means of support for teachers in reform mathematics education*. Paper presented at the annual meeting of the American Educational Research Association, Las Vegas, NV.
- Greenes, C., Ginsburg, H. P., & Balfanz, R. (2004). Big math for little kids. *Early Childhood Research Quarterly*, 19, 159–166. doi:10.1016/j.ecresq.2004.01.010
- Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education*, 2, 1–49.
- Griffin, S. A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 25–49). Cambridge, MA: MIT Press.
- Hatano, G. (2003). Forward. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. xi–xiii). Mahwah, NJ: Erlbaum.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- James, W. (1958). *Talks to teachers on psychology and to students on some of life's ideals*. New York, NY: Norton.
- Jordan, N. C., Hanich, L. B., & Uberti, H. Z. (2003). Mathematical thinking and learning difficulties. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Recent research and theory* (pp. 359–383). Mahwah, NJ: Erlbaum.
- Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, 22, 36–46. doi:10.1111/j.1540-5826.2007.00229.x
- Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, 45, 850–867. doi:10.1037/a0014939
- Jordan, N. C., & Levine, S. C. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Developmental Disabilities Research Reviews*, 15, 60–68. doi:10.1002/ddrr.46
- Krajewski, K. (2008). Prävention der Rechenschwäche [The early prevention of math problems]. In W. Schneider & M. Hasselhorn (Eds.), *Handbuch der Pädagogischen Psychologie* (pp. 360–370). Göttingen, Germany: Hogrefe.
- Krajewski, K., & Schneider, W. (2009). Early development of quantity to number–word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, 19, 513–526. doi:10.1016/j.learninstruc.2008.10.002
- Kroesbergen, E. H., Van Luit, J. E. H., Van Lieshout, E. C. D. M., Van Loosbroek, E., & Van de Rijt, B. A. M. (2009). Individual differences in early numeracy: The role of executive functions and subitizing. *Journal of Psychoeducational Assessment*, 27, 226–236. doi:10.1177/0734282908330586
- LeFevre, J., Fast, L., Skwarchuk, S., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, 81, 1753–1767. doi:10.1111/j.1467-8624.2010.01508.x
- Lembke, E., & Foegen, A. (2006, February). *Monitoring student progress in early math*. Paper presented at the 14th annual meeting of the Pacific Coast Research Conference, San Diego, CA.
- Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research & Practice*, 24, 12–20. doi:10.1111/j.1540-5826.2008.01273.x
- Leong, C. K., & Jerred, W. D. (2001). Effects of consistency and adequacy of language information on understanding elementary mathematics word problems. *Annals of Dyslexia*, 51, 275–298. doi:10.1007/s11881-001-0014-1
- Libertus, M. E., Feigenson, L. H., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, 14, 1292–1300. doi:10.1111/j.1467-7687.2011.01080.x
- Lonigan, C. J., Schatschneider, C., & Westberg, L. (2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. In *Developing early literacy: Report of the National Early Literacy Panel* (pp. 55–106). Washington, DC: National Institute for Literacy.
- Martinez, R. S., Missall, K. N., Graney, S. B., Aricak, O. T., & Clarke, B. (2008). Technical adequacy of early numeracy curriculum-based measurement in kindergarten. *Assessment for Effective Intervention*, 34, 116–125. doi:10.1177/1534508408326204
- Mazzocco, M., & Thompson, R. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice*, 20, 142–155. doi:10.1111/j.1540-5826.2005.00129.x
- Mix, K. S. (2009). How Spencer made number: First uses of the number words. *Journal of Experimental Child Psychology*, 102, 427–444. doi:10.1016/j.jecp.2008.11.003
- Morris, R. D., Stuebing, K. K., Fletcher, J. M., Shaywitz, S. E., Lyon, G. R., Shankweiler, D. P., . . . Shaywitz, B. A. (1998). Subtypes of reading disability: Variability around a phonological core. *Journal of Educational Psychology*, 90, 347–373. doi:10.1037/0022-0663.90.3.347
- Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 371–383). New York, NY: Routledge.
- Muldoon, K., Lewis, C., & Freeman, N. (2009). Why set-comparison is vital in early number learning. *Trends in Cognitive Sciences*, 13, 203–208. doi:10.1016/j.tics.2009.01.010
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through Grade 8 mathematics*. Reston, VA: Author.
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U. S. Department of Education.
- Palmer, A., & Baroody, A. J. (2011). Blake's development of the number words “one,” “two,” and “three”. *Cognition and Instruction*, 29, 265–296. doi:10.1080/07370008.2011.583370
- Piaget, J. (1964). Development and learning. In R. E. Ripple & V. N. Rockcastle (Eds.), *Piaget rediscovered* (pp. 7–20). Ithaca, NY: Cornell University.
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments & Computers*, 36, 717–731. doi:10.3758/BF03206553

- Purpura, D. J. (2010). *Informal number-related mathematics skills: An examination of the structure of and relations between these skills in preschool*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. AAT 3462344)
- Purpura, D. J., Hume, L., Sims, D., & Lonigan, C. J. (2011). Emergent literacy and mathematics: The value of including emergent literacy skills in the prediction of mathematics development. *Journal of Experimental Child Psychology*, 110, 647–658. doi:10.1016/j.jecp.2011.07.004
- Purpura, D. J., & Lonigan, C. J. (2013). Informal numeracy skills: The structure and relations among numbering, relations, and arithmetic operations in preschool. *American Educational Research Journal*, 50, 178–209. doi:10.3102/0002831212465332
- Rousselle, L., & Noel, M. (2007). Basic numerical skills in children with mathematics learning disabilities: A comparison of symbolic vs. non-symbolic number magnitude. *Cognition*, 102, 361–395. doi:10.1016/j.cognition.2006.01.005
- Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York, NY: Routledge.
- Sarnecka, B. W., & Carey, S. (2008). How counting presents number: What children must learn and when they learn it. *Cognition*, 108, 662–674. doi:10.1016/j.cognition.2008.05.007
- Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science*, 11, 655–661. doi:10.1111/j.1467-7687.2008.00714.x
- Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology*, 101, 545–560. doi:10.1037/a0014239
- Sinclair, A., Sigrist, F., & Sinclair, H. (1983). Young children's ideas about the written number system. In D. Rogers & J. A. Sloboda (Eds.), *The acquisition of symbolic skills* (pp. 535–541). New York, NY: Plenum Press.
- Song, M. J., & Ginsburg, H. P. (1987). The development of informal and formal mathematical thinking in Korean and U.S. children. *Child Development*, 58, 1286–1296.
- Stanovich, K. E., Siegal, L. S., & Gottardo, A. (1997). Converging evidence for phonological and surface subtypes of reading disability. *Journal of Educational Psychology*, 89, 114–127. doi:10.1037/0022-0663.89.1.114
- Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly*, 19, 99–120. doi:10.1016/j.jecresq.2004.01.002
- Stock, P., Desoete, A., & Roeyers, H. (2007). Early markers of arithmetic difficulties. *Educational and Child Psychology*, 24, 28–39.
- Swanson, H. L. (2004). Working memory and phonological processing as predictors of children's mathematical problem solving at different ages. *Memory & Cognition*, 32, 648–661. doi:10.3758/BF03195856
- Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology*, 96, 471–491. doi:10.1037/0022-0663.96.3.471
- Swanson, H. L., & Kim, K. (2007). Working memory, short-term memory, and naming speed as predictors of children's mathematical performance. *Intelligence*, 35, 151–168. doi:10.1016/j.intell.2006.07.001
- VanDerHeyden, A. M., Broussard, C., & Cooley, A. (2006). Further development of measures of early math performance for preschoolers. *Journal of School Psychology*, 44, 533–553. doi:10.1016/j.jsp.2006.07.003
- Woodcock, R., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson Tests of Achievement* (3rd ed.). Itasca, IL: Riverside.
- Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37, 197–206. doi:10.1086/651257

Received April 10, 2012

Revision received December 5, 2012

Accepted December 10, 2012 ■

Early Teacher Expectations Disproportionately Affect Poor Children's High School Performance

Nicole S. Sorhagen
Temple University

This research used prospective longitudinal data to examine the associations between first-grade teachers' over- and underestimation of their students' math abilities, basic reading abilities, and language skills and the students' high school academic performance, with special attention to the subject area and moderating effects of student demographic characteristics. Teachers' inaccurate expectations in first grade predicted students' math, reading comprehension, vocabulary knowledge, and verbal reasoning standardized test scores at age 15. Significant interactions between students' family income and teachers' misperceptions of students' math and language skills were found, such that teachers' over- and underestimation of abilities had a stronger impact on students from lower income families than on students from more affluent homes. In contrast, the effects of teachers' misperceptions of students' basic reading abilities on performance at age 15 did not differ by income. These results have implications for understanding the complexities of self-fulfilling prophecies and for understanding the achievement gap between students from disadvantaged and advantaged homes.

Keywords: self-fulfilling prophecy, teacher expectations, achievement

Some believe their academic successes can be, in part, attributed to a teacher who believed in their abilities; others can remember a teacher who made them doubt their intelligence, possibly instilling in them a lasting ambivalence toward learning. These reminiscences are both examples of how teacher–student relationships can inform and affect academic performance and cognitive development far into the future. Merton (1948) called attention to situations in which one person's inaccurate expectations and misperceptions about a second person lead the second person to act in ways that are similar to the first person's false beliefs, labeling them “self-fulfilling prophecies.” How long can such prophecies last, and are students from some backgrounds more affected by these self-fulfilling prophecies than other students?

In their classic study *Pygmalion in the Classroom*, Rosenthal and Jacobson (1968) demonstrated that schoolchildren's intellectual development can be shaped by teachers' expectations. The study created a heated debate among social scientists and educators about the validity of the Pygmalion experiment, the accuracy of teacher expectations, and the mechanisms by which self-fulfilling prophecies unfold in the classroom (for review, see Jussim & Harber, 2005; Spitz, 1999; Weinstein, 2002). Over the past 40 years, many investigators have studied the relation between a teacher's expectations and a student's academic performance and achievement using both experimental and naturalistic methods. Meta-analyses have provided evidence that experimentally in-

duced positive expectations can increase student performance (Raudenbush, 1984; Rosenthal, 1994; Rosenthal & Rubin, 1978; M. L. Smith, 1980). Naturalistic studies have demonstrated that children whose teachers underestimate their abilities achieve less in school than would be predicted on the basis of their test scores, whereas those whose abilities are overestimated achieve more (for review, see Jussim & Eccles, 1995; Jussim, Robustelli, & Cain, 2009; Weinstein, 2002). Although most studies have examined the short-term effects of teacher expectations on student achievement, some have shown that the effects sometimes persist for several years, even when students have changed teachers (Alvidrez & Weinstein, 1999; Hinnant, O'Brien, & Ghazarian, 2009; Rist, 1970; A. E. Smith, Jussim, & Eccles, 1999).

By and large, the magnitude of self-fulfilling prophecies in the classroom tends to be modest in experimental studies and to be smaller in naturalistic studies (Jussim et al., 2009; Rosenthal & Rubin, 1978). There is evidence, however, that teacher expectations have a more substantial impact on more vulnerable students, including students from low-income families, as well as low-achieving students, students who perceive differential treatment from teachers, and minority students (Brattesani, Weinstein, & Marshall, 1984; Jussim, Eccles, & Madon, 1996; Kuklinski & Weinstein, 2001; Madon, Jussim, & Eccles, 1997; McKown & Weinstein, 2008). In view of the long-standing and seemingly intractable achievement gap between American students from disadvantaged and advantaged homes, it is important to ask whether low expectations on the part of teachers have a lasting impact on the academic careers of children from lower income families.

Following one urban kindergarten class through second grade, Rist (1970) found that preferential treatment was given to children from higher social class families in the form of seating assignments, and, as a result, these children received more teacher attention. As the children progressed through first and second

This article was published Online First March 25, 2013.

Special thanks are extended to Laurence Steinberg for his contributions on an earlier version of this article and Marsha Weinraub for her helpful comments and assistance.

Correspondence concerning this article should be addressed to Nicole S. Sorhagen, Department of Psychology, Temple University, 1701 North 13th Street, Philadelphia, PA 19122. E-mail: nicole.sorhagen@temple.edu

grade, fewer children from the low social class families were able to "move up" to the more esteemed table of the teacher. More recent investigations have revealed stronger self-fulfilling effects on lower income students' math achievement in third grade, compared to more wealthy students, based on teachers' inaccurate expectations of math abilities in first grade (Hinnant et al., 2009), and on lower income students' school performance in sixth grade and math achievement in seventh grade based on teachers' perceptions of math effort in sixth grade (Jussim et al., 1996).

There is little evidence that the reading abilities of students from lower income families are more affected by teacher expectations, though there is evidence for moderation in other vulnerable groups of children. First-grade teachers' over- and underestimation of minority boys' reading abilities predict reading achievement in third grade (Hinnant et al., 2009). Teachers' expectations of reading abilities have also been found to influence later reading abilities in classrooms where students perceive differential treatment from teachers (Brattesani et al., 1984; McKown & Weinstein, 2008). Researchers have also considered the effect of tracking and grouping students by reading abilities because teachers choose these groups. Students placed in advanced reading groups tend to perform better on measures of reading achievement regardless of their prior reading abilities, compared to students placed in low-ability groups (Eder, 1981; Weinstein, 1979).

Teacher's expectations in academic subject areas other than math and reading has not been well explored. Certainly, early abilities influence later abilities. For example, vocabulary size is a strong predictor of later academic performance, especially reading ability (Baumann, Kame'enui, & Ash, 2003; Stanovich, 1986). In fact, there seems to be a Matthew effect ("the rich get richer and the poor get poorer") in word learning such that the more words children know, the more easily they seem to learn additional words (Stanovich, 1986, 2000). Moreover, converging evidence suggests that early phonological awareness is a substantial factor in the development of basic and advanced reading comprehension (Blachman, 2000; Bus & van IJzendoorn, 1999; Farkas & Beron, 2004; Johnston, Anderson, & Holligan, 1996; NICHD Early Child Care Research Network, 2005b; Stuart & Coltheart, 1988; Wagner et al., 1997).

Social class differences in children's vocabulary size have been found as early as age 3 such that young children from high socioeconomic status (SES) families tend to know more words than children from low-SES families, apparently as a result of exposure to more words starting from infancy (Hart & Risley, 1995; Schachter, 1979). Additionally, children from low-SES families tend to have less phonological understanding and spoken-language abilities than children from more affluent SES families (Farkas & Beron, 2004; Hecht, Burgess, Torgesen, Wagner, & Rashotte, 2000; Locke, Ginsborg, & Peers, 2002). These early differences in vocabulary and phonological awareness are especially problematic given that the gap in knowledge between low- and high-income students seems to increase with age (Lee & Burkam, 2002).

What if the type of academic abilities that teachers over- and underestimate matters? Do misperceptions of different academic subjects affect students from different demographic backgrounds differently? This study used data from the NICHD Study of Early Child Care and Youth Development (SECCYD) to examine whether teacher expectations in the first year of elementary school

continue to influence student achievement up to age 15 and whether these effects might be especially pronounced among lower income students. The NICHD SECCYD is ideally suited for the examination of long-term teacher expectancy effects because it tracked children's academic experiences from elementary school through high school, considered a wide range of children's academic abilities, and collected both standardized achievement test scores and teacher assessments of students' abilities at multiple points in time.

Hinnant and colleagues (2009) also used data from the NICHD SECCYD but did not consider basic reading achievement separately from vocabulary knowledge and comprehension. In light of the vulnerability of students from low-income families in vocabulary and phonological awareness, the present study considered early language skills separately from early basic reading abilities. It is possible that high early teacher expectations create the exposure to new words that young students from lower income families need to catch up to their more affluent peers, who presumably have larger vocabularies at school entry. In addition, Hinnant and colleagues considered the effects of teacher expectations through fifth grade; the present study extended this line of investigation by considering the effects of teacher expectations through high school.

Based on prior work with this and other samples, mismatches between teachers' estimations of a child's academic ability and a child's actual academic achievement in early elementary school were hypothesized to have small but lasting effects on academic achievement in high school math, reading comprehension, word knowledge, and verbal reasoning ability. Inaccurate teacher expectations were anticipated to have a larger effect on students from lower income families than on more affluent students. Furthermore, under- and overestimation of language skills were anticipated to be more important for students from lower income families compared to misperceptions of basic reading abilities. Comparable analyses were conducted to examine the interaction between teacher expectations and students' gender and race.

Following the practice of previous researchers, teacher expectations were operationalized by computing a discrepancy score between a first-grade teacher's report of a child's academic performance and a child's performance on standardized tests while in first grade (Hinnant et al., 2009; Jussim & Harber, 2005; Madon et al., 1997). Because teacher expectations may be influenced by nonacademic factors, such as a child's attentiveness, self-reliance, and classroom behavior, the present research controlled for first-grade classroom observer ratings of child competence and maternal reports of a child's self-control. This allowed for the consideration of known variables that affect student performance and that likely influence teacher expectations but are not assessed through standardized achievement tests.

Method

Participants

Participants included American children recruited at birth and followed through age 15 who were enrolled in the NICHD SECCYD. New mothers were recruited from 24 hospitals in 10 data collection sites in 1991 (Little Rock, Arkansas; Irvine, Cali-

fornia; Lawrence, Kansas; Boston, Massachusetts; Philadelphia, Pennsylvania; Pittsburgh, Pennsylvania; Charlottesville, Virginia; Morganton, North Carolina; Seattle, Washington; and Madison, Wisconsin). A total of 8,986 mothers were screened for eligibility at the hospital within 48 hours after birth. Mother–infant dyads were excluded from the study if the mother was younger than 18 years old, could not speak English, was not healthy, or refused; if the infant had a serious medical condition; or if the family planned to move out of the area within a year or lived in a dangerous neighborhood. Based on the criteria, the sample was reduced to 5,416 mother–infant dyads. A conditional random sampling plan was used to select 3,015 mother–infant dyads. Of that group, 1,364 families were successfully recruited and completed the 1-month interview (families were excluded if they refused, could not be contacted, planned to move from the area within the next 3 years, or if the infant was hospitalized for more than 7 days). The final sample was socioeconomically and ethnically diverse, including 24% ethnic-minority children, 10% mothers with less than a high school education, and 14% single parents. Although the sample is not nationally representative, it is one of the largest and richest longitudinal studies of American schoolchildren ever conducted. Assessments were conducted when the children were 6, 15, 24, 36, and 54 months old, while in the first, third, and fifth grades, and at age 15, with individual standardized tests, observations of families and school settings, and parent and teacher reports of behavior (further details can be found at NICHD Early Child Care Research Network, 2005a, and <https://secc.rti.org>).

For the present study, the sample was restricted to White and African American students. There is reason to think, based on the literature on stereotypes, prejudice, and cross-cultural psychology, that self-fulfilling prophecies in the classroom may affect students of different ethnicities differently. The small number of Asian, American Indian, and “other” students (approximately 6% of the total sample) made it impractical to keep them in the analyses. This brought the sample to 1,273 students.

Within the restricted sample, 894 first-grade teachers participated in the study. Most teachers had one study child in their classroom (53 teachers had two study children, 6 teachers had three, and 1 had eight). The teachers were mostly female (96%), and 94% were White. They had an average of 14.36 years ($SD = 9.25$) of overall teaching experience and 9.05 years ($SD = 8.03$) of teaching first grade. The classrooms had an average of 21 students ($SD = 5.30$), and most of the classrooms had a majority of White students (75%, $SD = 27$). Sixty-seven percent of the students were grouped for reading instruction, and 23% were grouped for math instruction. Most schools were public (80%) and began at kindergarten (63%); many ended in fifth grade (41%).

Due to the longitudinal nature of the data there were some missing values in the variables used in the present study.¹ Students with missing data were included in the analyses in order to avoid bias. Table 1 presents descriptive statistics of the 967 students who had valid teacher-related variables because teacher reports contributed to the outcome variables of the first set of analyses and to the main variables of interest in the second set of analyses. Relations between the variables are shown in Table 2. It is important to note that the sample size varied between each model presented below.

Table 1
Descriptive Statistics

Variable	N (%)	M	SD
Demographics			
Gender (female)	967 (49%)		
Ethnicity (African American)	967 (12%)		
Family income-to-needs ratio	959	3.96	3.05
Child competence	910	8.98	1.58
Self-control	962	13.04	3.35
Discrepancy scores			
Math	905	−0.01	0.80
Basic reading	906	0.00	0.73
Language skills	900	0.00	0.89
Achievement scores (WJ-R)			
Letter-Word, 54 months	928	369.59	20.99
Letter-Word, first grade	928	452.74	23.85
Picture Vocabulary, 54 months	931	460.12	14.02
Picture Vocabulary, first grade	924	486.14	8.77
Picture Vocabulary, 15 years	771	518.82	13.14
Verbal Analogies, 15 years	773	525.85	14.16
Passage Comprehension, 15 years	770	520.56	12.68
Applied Problems, 54 months	926	425.15	19.26
Applied Problems, first grade	928	470.30	15.47
Applied Problems, 15 years	769	524.77	16.75
Academic skills questionnaire			
Language literacy	943	3.35	0.96
Mathematical thinking	941	3.21	0.95

Note. WJ-R = Woodcock-Johnson—Revised Test of Achievement and Cognitive Abilities.

Measures

Demographics. During home interviews when children were 1 month old, mothers reported demographic information, including the child’s gender and ethnicity (African American, European American, Asian, American Indian, or other). In the first-grade interview, mothers provided information on family income. Family income-to-needs ratios were created by dividing the poverty threshold for the household size by the reported family income. When available, missing data were imputed by averaging the family income-to-needs ratio reported at kindergarten and third grade. For the present study, 37% of families had an income-to-needs ratio below 2.5, 38% had an income-to-needs ratio between 2.5 and 4, and 25% had a ratio over 5.

¹ Dichotomous dummy codes were created to indicate missing values in order to determine if attrition in the outcomes in high school varied as a function of the covariates and predictors. Separate logistic regressions for the four outcome variables were then run predicting the missing values from the covariates that were included in the main analysis. The basic reading and the language skills discrepancy scores were both entered into the same model for each of the related outcomes here. The results indicated that the overall model was not significant for the WJ-R Applied Problems, Passage Comprehension, Picture Vocabulary, and Verbal Analogies scores: $\chi^2(7, n = 825) = 8.58, ns$; $\chi^2(8, n = 821) = 10.94, ns$; $\chi^2(8, n = 824) = 9.59, ns$; and $\chi^2(8, n = 824) = 10.56, ns$; respectively). *T* tests with Bonferroni adjustment indicated that many of the individual predictors were not significantly related to missing values—with the exception that students with missing Applied Problems scores in high school tended to score lower on prior math abilities ($M = 420.93, SD = 21.35$) compared to students without missing data ($M = 425.87, SD = 18.50$), $t(298.93) = 3.06, p < .01$.

Table 2
Correlations Among Predictor and Outcome Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. WJ-R Applied Problems, 54 months	1														
2. WJ-R Letter-Word, 54 months	.557***	1													
3. WJ-R Picture Vocabulary, 54 months	.543***	.481***	1												
4. Child competence	.196***	.131***	.049	1											
5. Self control	.161***	.141***	.178***	.071*	1										
6. Gender (female)	.130***	.097**	-.069*	.064	.069*	1									
7. Ethnicity (African American)	-.351***	-.203***	-.356***	-.060	-.141***	-.018	1								
8. Family income-to-needs ratio	.304***	.314***	.334***	.062	.201***	.040	-.264***	1							
9. Math discrepancy score	.059	.145***	-.001	.184***	.075*	.005	-.037	.022	1						
10. Language skills discrepancy score	.280***	.307***	.044	.275***	.113**	.139***	-.057	.083*	.606***	1					
11. Basic reading discrepancy score	.195***	.140***	.078*	.182***	.097**	.069*	-.049	.028	.579***	.786***	1				
12. WJ-R Applied Problems, 15 years	.495***	.461***	.431***	.163***	.178***	-.100**	-.306***	.315***	.130***	.268***	.176***	1			
13. WJ-R Passage Comprehension, 15 years	.511***	.453***	.503***	.052	.165***	.046	-.351***	.307***	.106**	.245***	.167***	.684***	1		
14. WJ-R Picture Vocabulary, 15 years	.442***	.431***	.601***	.048	.152***	-.098**	-.402***	.296***	.080*	.128**	.127**	.613***	.714***	1	
15. WJ-R Verbal Analogies, 15 years	.537***	.483***	.454***	.130***	.208***	.006	-.354***	.303***	.168***	.295***	.223***	.692***	.704***	.635***	1

Note. WJ-R = Woodcock-Johnson—Revised Test of Achievement and Cognitive Abilities.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Standardized assessment of achievement. The children in the sample were administered the Woodcock-Johnson—Revised (WJ-R) Test of Achievement (WJ-R ACH) and Test of Cognitive Abilities (WJ-R COG) multiple times throughout the study (Woodcock & Johnson, 1989). Based on normative data, the WJ-R has good reliability (McGrew, Werder, & Woodcock, 1991). Internal consistency ranged from the high .80s to the .90s. Test-retest reliability ranged from the .60s to the .80s. The WJ-R also has excellent predictive validity across the life span (McGrew, 1993; McGrew & Hessler, 1995; McGrew & Knopik, 1993) and is highly correlated with other tests of cognitive abilities and achievement (McGrew et al., 1991). The *W* scores of the WJ-R were used for the present analyses. *W* scores are special transformations of the Rasch ability scale converted from the raw scores. They are centered at a value of 500 to allow for comparisons across standardized tests and ages, making it possible to assess individual development over time.

Three subscales of the WJ-R ACH were used to identify math, basic reading, and advanced reading abilities. The Applied Problems subscale measured math abilities and required children to analyze and solve practical word and story problems with math calculations. The Letter-Word subscale measured basic reading abilities and required children to identify letters and words. The Passage Comprehension subscale measured advanced reading abilities and required children to read a short passage and identify a missing key word or to match a picture to a phrase. Three subscales of the WJ-R COG were used to measure vocabulary size, phonological awareness, and verbal reasoning abilities. The Picture Vocabulary subscale measured language development and vocabulary comprehension and required children to recognize or name pictures of objects. The Incomplete Words subscale measured phonological awareness and required children to identify incomplete spoken words. Finally, the Verbal Analogies subscale measured verbal reasoning and required children to complete phrases with appropriate analogies.

The Letter-Word, Picture Vocabulary, and Applied Problems subscales administered at 54 months were used as controls for previous academic ability, that is, before first grade. Subscales of the WJ-R administered in first grade (Applied Problems for math, Letter-Word for reading, Incomplete Words and Picture Vocabulary for language skills) served as indicators of academic achievement in that year and were used to compute the teacher discrepancy scores for each child (see below). Finally, scores on the Passage Comprehension, Picture Vocabulary, Verbal Analogies, and Applied Problems WJ-R subscales administered at 15 years were the outcome measures.

Teacher rating of student ability. The academic skills questionnaire was a teacher report of student math, reading, and language abilities. First-grade teachers were instructed to rate the study child's academic skills and performance compared to other children at the same grade level on a 5-point scale (1 = *not yet*, 5 = *proficient*). The language literacy subscale asked teachers to rate students' skills related to listening, speaking, and early reading and writing behaviors. Example items are "This child reads first grade books independently with comprehension, for example, reads most words correctly and answers questions about what was read, makes predictions while reading, and retells story after reading" and "This child reads words with irregular vowels sounds, for example, reads 'through', 'point', 'enough', or 'shower'." The mathematical

thinking subscale asked about a student's ability to perceive, understand, and use skills in solving math problems and math-related activities. An example item is "Demonstrates an understanding of place values, for example, by explaining that fourteen is ten plus four, or using two stacks of ten and five single cubes to represent the number 25." Both measures had high internal reliability (language literacy subscale, Cronbach's $\alpha = .96$; mathematical thinking subscale, Cronbach's $\alpha = .94$).

Teacher discrepancy score. Teacher's under- and overestimation of a student's abilities was operationalized by computing a discrepancy score between the first-grade teacher's report of a child's academic performance and the first-grader's performance on standardized tests. The teachers' ratings of students' abilities and the students' WJ-R scores in first grade were related (mathematical thinking scale and Applied Problems, $r^2 = .531$; language literacy scale and Incomplete Words, $r^2 = .286$, Letter-Word, $r^2 = .642$, and Picture Vocabulary, $r^2 = .230$; all $ps < .001$). Teacher discrepancy scores were computed by regressing teacher perceptions of a child's math ability (the mathematical thinking scale) on the child's math abilities (WJ-R Applied Problems), and teacher perceptions of the student's language literacy ability (the language literacy scale) on the student's basic reading abilities (WJ-R Letter-Word) and language skills (WJ-R Picture Vocabulary and Incomplete Words).² The resulting residual scores (math, basic reading, and language skills) provide an index of the extent to which teacher perceptions of ability vary from a child's observed performance; that is, they are a measure of a teacher's rating of a child's ability with the child's standardized ability removed (Hinnant et al., 2009; Jussim & Harber, 2005; Madon et al., 1997).

A negative residual score reflects teacher underestimation, and a positive residual score reflects teacher overestimation. The closer a residual score is to zero, the smaller the discrepancy between teacher expectations and actual performance. Unstandardized residuals were used. The discrepancy scores for math ranged from -2.45 to 1.96 , those for reading ranged from -3.13 to 2.42 , and those for language skills ranged from -2.41 to 1.91 . The distribution of scores suggests that teachers in the present study were generally accurate. For all of the discrepancy scores, approximately 64% of the teachers were within one standard deviation from the mean and approximately 18% were above or below one standard deviation from the mean.

Noncognitive covariates. As noted earlier, teacher expectations may be influenced not only by their perceptions of a student's intellectual abilities but by a host of noncognitive factors. Because these factors may also influence student achievement, it was necessary to take these variables into account in the analyses. Two first-grade noncognitive variables were included in the analyses: maternal report of her child's self-control and independent observer ratings of student competence. Mother's assessment of her child's self-control was measured by the Self-Control subscale of the Social Skills Rating System (SSRS, Cronbach's $\alpha = 0.88$). An example item is "ends disagreements with you calmly." Observer ratings of children's competence were collected using the Classroom Observer System. Trained observers recorded and rated a child's behavior in the classroom. The child competence composite was created by summing observer ratings of the child's self-reliance and positive affect. Higher scores indicate a higher degree of child competence. The present analyses were also conducted with other similar covariates derived from classroom observations,

maternal reports, tests of planning, and inhibitory control. When these alternative cognitive and noncognitive covariates were controlled for, the significance of the main effects and interactions predicting age 15 achievement reported below remained unchanged.

Analytic Plan

Data management followed the guidelines of Tabachnick and Fidell (2007).³ The income-to-needs variable was centered on its mean to avoid problems of collinearity and to facilitate interpretation. Ethnicity was converted into a dichotomous dummy variable comparing White students to African American students.

Preliminary analyses were first conducted to investigate which, if any, demographic, cognitive and noncognitive covariates predicted each of the teacher discrepancy scores. Income by ethnicity and gender by ethnicity interaction terms were also entered into the models to assess bias of teachers' misperceptions of abilities within potentially stigmatized groups.

In order to test the central hypotheses, a series of hierarchical multiple regression analyses were used to predict the WJ-R scores at age 15 from the teacher discrepancy scores, after controlling for gender, ethnicity, family income, and the relevant test scores at 54 months, as well as observer ratings of child competence and maternal ratings of self-control. In order to examine whether teacher expectations differentially affected children from different demographic groups, the interactions between the teacher discrepancy score and gender, ethnicity, and family income were examined in the prediction of age 15 achievement, after controlling for the main effects of demographic variables, teacher discrepancy

² A factor analysis was done on the language literacy academic skills questionnaire in an attempt to parse out language skills from reading skills. A principal components extraction was used to estimate the number of factors and to assess the favorability of the correlation matrices. Two factors were extracted. When orthogonal varimax rotation with principal factor extraction was performed, the interpretation of the two factors was not ideal for this analysis. For the most part, the questions on the students' reading and language skills loaded equally onto both factors, while questions on writing loaded on Factor 1.

The language literacy scale was also used to compute two new composites that reflect a teacher's perception of a student's reading abilities and his or her perception of a student's language skills. Four questions on reading abilities were used to create a reading scale, and four questions on vocabulary and phonological abilities were used to create a language skills score. Following the procedures outlined by the NICHD data report for the academic skills questionnaire, values of 6 (*n/a*) were recoded to 1 (*not yet*) if at least 60% of the possible responses had values from 1 to 5. If less than 60% of the responses had values from 1 to 5, then the values of 6 were coded as missing ($n = 10$). The new "reading only" and "language skill only" scores were computed as the mean of the four questions. Cronbach's alpha for each suggested high reliability for the sample used in the analysis (reading, Cronbach's $\alpha = .88$; language skills, Cronbach's $\alpha = .91$). These scores were then used to create the teacher discrepancy scores for each domain. Then the "reading only" and "language skills only" discrepancy scores were used in the same data analyses described in the Analytic Plan and Results sections of the present study. The findings followed a pattern similar to that when the discrepancy scores created from the entire language literacy scale were used, which included a teacher's perceptions of other abilities (e.g., writing).

³ Univariate and multivariate outliers were assessed for each model. One case with extremely low z scores on the WJ-R scores at 54 months and 15 years was a univariate outlier well beyond the $p = .001$ cut-off (over 5) and was deleted.

scores, WJ-R scores at 4.5 years, and noncognitive covariates. Three-way interaction terms were then entered into the models to examine the effect of the teacher discrepancy scores on high school achievement for students from multiple vulnerable groups. Teacher discrepancy score by ethnicity by gender and teacher discrepancy score by income by gender interaction terms were entered, controlling for demographic and noncognitive covariates, teacher discrepancy scores, and all subsequent two-way interactions. Seven separate models were run: one predicting math and two predicting each advanced reading, vocabulary knowledge, and verbal reasoning, one with the basic reading discrepancy score as a predictor and the other with the language skills discrepancy score as a predictor.

In the present sample, family income differed between White and African American students, $t(349.64) = 15.34, p < .001$: for White students, $M = 4.26, SD = 3.22$, range = 0.10–25.05; for African American students, $M = 1.73, SD = 1.44$, range = 0.11–7.57. This disparity could confound the results of models containing the income by ethnicity two-way interaction term and the teacher discrepancy score by income by ethnicity three-way interaction term. In order to investigate these interactions, analyses of all models were rerun with a restricted sample of students with family income-to-needs ratios of 2.25 or below, based on the work of Burchinal and colleagues (2011), who selected students with family incomes of 225% of the poverty threshold or below from the NICHD data set to investigate the ethnicity achievement gap in America. In the present study, the restricted sample included 80% of African American students and 31% of White students.

Results

Preliminary Analyses

Preliminary regression analyses were conducted to assess whether child demographic, cognitive, and noncognitive characteristics predicted teacher misperceptions of students' abilities (see Table 3). Step 1 of the regression analyses showed that students with higher scores on the child competence measure tended to have higher math, basic reading, and language skills discrepancy scores ($\beta = .178, p < .001$, $\beta = .160, p < .001$, and $\beta = .258, p < .001$, respectively).

In addition, WJ-R Letter-Word scores at 4.5 years and self-control predicted teacher's misperceptions of reading abilities ($\beta = .113, p < .001$ and $\beta = .072, p < .05$, respectively). Finally, language skills discrepancy scores were higher for female students and students with higher self-control ($\beta = .116, p < .001$ and $\beta = .076, p < .05$, respectively).

There was no evidence that teachers' misperceptions of abilities were influenced by a student's membership in multiple stigmatized groups because the ethnicity by income and ethnicity by gender interaction terms did not significantly predict discrepancy scores for any of the models when entered into Step 2. Since child competence predicted all of the teacher discrepancy scores, interaction terms between student demographic characteristics and child competence were also tested. None were significant.

Predicting Age 15 Achievement From First-Grade Teacher Expectations

When teachers underestimated student's abilities in the first grade, the student's WJ-R scores at age 15 were lower, even after taking into account prior measures of ability, gender, ethnicity, family income, and noncognitive factors known to influence achievement. On the other hand, when a student's academic abilities were overestimated, his or her later performance on the WJ-R was higher, again controlling for prior academic ability, demographics, and the noncognitive covariates (see Tables 4, 5, and 6).

Consistent with the hypotheses, the inclusion of the teacher discrepancy scores accounted for 6% or less of the total variance explained in all models (see ΔR^2 s of Step 2 in Tables 4, 5, and 6). Similarly, the standardized effects of teacher discrepancy scores predicting age 15 math, advanced reading, vocabulary knowledge, and verbal reasoning were small but significant (for math, $\beta = .084, p < .01$; for basic reading predicting Passage Comprehension, $\beta = .105, p < .01$, Picture Vocabulary, $\beta = .079, p < .05$, and Verbal Analogies, $\beta = .166, p < .001$; and for language skills predicting Passage Comprehension, $\beta = .131, p < .001$, Picture Vocabulary, $\beta = .103, p < .01$, and Verbal Analogies, $\beta = .247, p < .001$).

Curvilinear effects of teachers' inaccurate perceptions of abilities were tested by entering the squared teacher discrepancy scores

Table 3
Regression Analyses Predicting Teacher Discrepancy Scores

Predictor	Math discrepancy score ^a			Language skills discrepancy score ^b			Reading discrepancy score ^c		
	B	SE B	β	B	SE B	β	B	SE B	β
Step 1	-1.205	0.654		-1.751	1.058		-2.330	0.463	
Gender (female)	-0.018	0.054	-.011	0.206	0.058	.116***	0.063	0.049	.044
Ethnicity (African American)	-0.036	0.084	-.015	-0.049	0.090	-.019	-0.036	0.073	-.017
Family income-to-needs ratio	-0.002	0.009	-.009	0.012	0.010	.042	-0.009	0.008	-.038
Child competence	0.090	0.017	.178***	0.145	0.018	.258***	0.074	0.015	.160***
Self-control	0.014	0.008	.061	0.020	0.009	.076*	0.016	0.007	.072*
WJ-R Applied Problems, 4.5 years	0.001	0.002	.012						
WJ-R Picture Vocabulary, 4.5 years				0.000	0.002	.003			
WJ-R Letter-Word, 4.5 years							0.004	0.004	.113**

Note. WJ-R = Woodcock-Johnson—Revised Test of Achievement and Cognitive Abilities.

^a $F(6, 859) = 5.72, p < .001, R^2 = .038$. ^b $F(6, 854) = 15.90, p < .001, R^2 = .100$. ^c $F(6, 860) = 8.29, p < .001, R^2 = .055$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Regression Analyses Predicting WJ-R Applied Problems Scores

Predictor	WJ-R Applied Problems		
	B	SE B	β
Step 1	368.796	12.507	
WJ-R, 4.5 years	0.348	0.030	.401***
Child competence	0.773	0.330	.073*
Self-control	0.347	0.157	.070*
Gender (female)	-5.381	1.032	-.161***
Ethnicity (African American)	-5.508	1.600	-.114**
Family income-to-needs ratio	0.872	0.173	.166***
Step 2	370.829	12.477	
WJ-R, 4.5 years	0.347	0.030	.400***
Child competence	0.614	0.334	.058
Self-control	0.321	0.156	.065*
Gender (female)	-5.349	1.028	-.160***
Ethnicity (African American)	-5.445	1.593	-.113**
Family income-to-needs ratio	0.876	0.173	.166***
Math discrepancy score	1.770	0.650	.084**
Step 3	370.777	12.440	
WJ-R, 4.5 years	0.348	0.030	.401
Child competence	0.602	0.333	.057
Self-control	0.321	0.156	.064
Gender (female)	-5.392	1.025	-.162
Ethnicity (African American)	-5.317	1.589	-.110*
Family income-to-needs ratio	0.869	0.172	.165
Math discrepancy score	1.751	0.648	.083
Math Discrepancy Score \times Income	-0.516	0.226	-.069*

Note. Step 1: $F(6, 721) = 58.69, p < .001, R^2 = .328$; Step 2: $F(7, 720) = 51.81, p < .001, \Delta R^2 = .007^{**}$; Step 3: $F(8, 719) = 46.25, p < .001, \Delta R^2 = .005^*$. WJ-R = Woodcock-Johnson—Revised Test of Achievement and Cognitive Abilities.

* $p < .05$. ** $p < .01$. *** $p < .001$.

into the models after the teacher discrepancy scores. None of the terms was significant, so they were not included in the final model.

Moderators of Teacher Expectations

Next, moderation was tested by entering the interaction terms into each model after all lower order interaction terms, main effects, and controls. No three-way interaction terms were significant predictors of high school achievement. For models using the basic reading discrepancy score, interactions between the teacher discrepancy score and gender, ethnicity and family income did not predict age 15 achievement as a block, nor did the single interaction terms (i.e., there was no evidence for moderation found for the models shown in Table 5).

The block of two-way interaction terms also did not add to the prediction of high school achievement for models that used the math and language skills discrepancy scores as predictors, though the teacher discrepancy score by family income interaction terms were significant predictors in all models. The teacher discrepancy score by family income interaction terms were then entered separately. The subsequently added teacher discrepancy score by gender and teacher discrepancy score by ethnicity interaction terms were not significant predictors of high school achievement as a block or individually, and they were dropped from the final model. Step 3 in Tables 4 and 6 shows the results of the final models.

The addition of the teacher discrepancy score by family income interaction term accounted for a significant, but less than 1%,

portion of the variance in all models, which is small when measured against traditional effect size benchmarks (Cohen, Cohen, West, & Aiken, 2003). Aguinis, Beaty, Boik, and Pierce (2005), however, found that the median effect size of interaction terms (measured in Cohen's f^2) was .002 across 30 years in prominent applied psychology and management journals. In the present analyses, the Cohen's f^2 ($R_{AB}^2 - R_{AB}^2/1 - R_{AB}^2$) associated with the addition of the teacher discrepancy score by family income interaction term was as follows: .007 for Applied Problems, .009 for Passage Comprehension, .009 for Picture Vocabulary, and .006 for Verbal Analogies.

To illustrate the significant interactions, the relations between the family income-to-needs ratio and the age 15 WJ-R scores were plotted separately for middle (the mean), high (+1 *SD*), and low (-1 *SD*) teacher discrepancy scores (Aiken & West, 1991). All covariates were set to their sample means. Figure 1 shows that when teacher expectations are high, all students, regardless of family income, tend to perform well on the WJ-R subscales. However, the impact of underestimating math abilities and language skills is far worse for low-income students. The range of scores predicted by the teacher discrepancy scores accounts for 20 or fewer points on the WJ-R subscales. In terms of the present sample, the range of predicted scores for the WJ-R Applied Problems subscale (519 to 532) is equivalent to going from the 42nd percentile to the 73rd percentile. Likewise, going from a 516 to a 524 on the WJ-R Passage Comprehension subscale is similar to going from the bottom 36% to the top 63% of the sample. For the WJ-R Picture Vocabulary model, the range of predicted scores (from 515 to a 522) is equivalent to 21% and 36% of the sample, and for Verbal Analogies, the range (from 519 to 530) is equal to about 30% and 60%.

The significant interactions were probed with Hayes and Matthes (2009) SPSS Macro, which probes single DF interactions in ordinary least squares regression models using the same low, middle, and high teacher discrepancy score indicators as in the interaction plots described above.

When teachers underestimated students' abilities, each 1-point increase in the family income-to-needs-ratio was associated with a 0.80-point increase in advanced reading ($\beta = .201$, 95% CI [.15, .25]), a 0.67-point increase in vocabulary size ($\beta = .149$, 95% CI [.11, .19]), a 0.88-point increase in verbal reasoning ($\beta = .198$, 95% CI [.15, .25]), and a 1.24-point increase in math ($\beta = .235$, 95% CI [.19, .28]) measured by WJ-R scores in high school, holding all other variables constant.

Similarly, when teachers are accurate in their expectations, a 1-point increase in the family income-to-needs ratio is associated with a half-point increase in the WJ-R Passage Comprehension score ($\beta = .128$, 95% CI [.09, .16]), a 0.37-point increase in the WJ-R Picture Vocabulary score ($\beta = .083$, 95% CI [.05, .11]), a half-point increase in the WJ-R Verbal Analogies score ($\beta = .114$, 95% CI [.08, .15]), and a 0.74-point increase in the WJ-R Applied Problems score ($\beta = .151$, 95% CI [.12, .19]) in high school, holding all other variables constant.

Students whose early math abilities and language skills are overestimated perform the same regardless of their family income (for Passage Comprehension, $\beta = .055$, 95% CI [.01, .10]; for Picture Vocabulary, $\beta = .016$, 95% CI [-.03, .06]; for Verbal Analogies, $\beta = .03$, 95% CI [-.02, .08]; for math, ($\beta = .067$, 95% CI [.02, .12]).

Table 5

Regression Analyses Predicting WJ-R Passage Comprehension, Picture Vocabulary, and Verbal Analogies Scores Using Basic Reading Discrepancy Scores as Predictor

Predictor	WJ-R Passage Comprehension ^a			WJ-R Picture Vocabulary ^b			WJ-R Verbal Analogies ^c		
	B	SE B	β	B	SE B	β	B	SE B	β
Step 1	441.536	7.554		304.919	13.657		361.825	16.299	
WJ-R, 4.5 years	0.211	0.020	.355***	0.464	0.029	.501***	0.331	0.035	.331***
Child competence	-0.171	0.253	-.021	0.074	0.237	.009	0.786	0.283	.088**
Self-control	0.281	0.121	.075*	0.140	0.114	.036	0.422	0.137	.100**
Gender (female)	-0.22	0.793	-.001	-1.874	0.747	-.072*	0.403	0.891	.014
Ethnicity (African American)	-8.695	1.196	-.239***	-7.893	1.167	-.208***	-7.874	1.393	-.192***
Family income-to-needs ratio	0.486	0.135	.123***	0.254	0.127	.061*	0.527	0.152	.118**
Step 2	445.753	7.613		308.418	13.654		369.73	16.075	
WJ-R, 4.5 years	0.204	0.020	.343***	0.459	0.029	.496***	0.321	0.035	.320***
Child competence	-0.305	0.254	-.038	-0.038	0.239	-.005	0.532	0.282	.059
Self-control	0.253	0.120	.067*	0.118	0.114	.030	0.372	0.134	.088**
Gender (female)	-0.137	0.789	-.005	-1.993	0.745	-.076**	0.132	0.877	.005
Ethnicity (African American)	-8.630	1.188	-.237***	-7.854	1.162	-.207***	-7.785	1.368	-.190***
Family income-to-needs ratio	0.501	0.134	.126***	0.261	0.127	.063*	0.544	0.149	.122***
Basic reading discrepancy score	1.810	0.552	.105***	1.430	0.520	.079**	3.229	0.612	.166***

Note. WJ-R = Woodcock-Johnson—Revised Test of Achievement and Cognitive Abilities.

^a Step 1: $F(6, 721) = 50.15, p < .001, R^2 = .294$; Step 2: $F(7, 720) = 45.10, p < .001, \Delta R^2 = .010^{***}$. ^b Step 1: $F(6, 722) = 87.26, p < .001, R^2 = .420$; Step 2: $F(7, 721) = 76.56, p < .001, \Delta R^2 = .006^{***}$. ^c Step 1: $F(6, 724) = 49.62, p < .001, R^2 = .291$; Step 2: $F(7, 723) = 48.09, p < .001, \Delta R^2 = .026^{***}$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Additional simple slope analyses were then conducted in order to assess the magnitude of the self-fulfilling effects across different academic subjects and student's economic background by considering the moderating influence of income on the effect of the teacher discrepancy scores predicting high school achievement. Furthermore, the Johnson-Neyman technique, which finds the exact value of a continuous moderator where the relation between two variables becomes insignificant, was used to find the specific point on the income-to-needs ratio where the discrepancy scores no longer predicted later WJ-R scores (Hayes & Matthes, 2009). The vertical lines in Figure 1 represent these points.

The results indicated that, holding all other variables constant, the language skills discrepancy score was significantly related to Passage Comprehension, Picture Vocabulary, and Verbal Analogies up to income-to-needs ratios of 6.22, 5.96, and 9.03, respectively. Holding all other variables constant, the self-fulfilling effects were stronger at the poverty threshold (an income-to-needs ratio of 1) for all models (for Passage Comprehension, $\beta = .201$, 95% CI [.16, .25]; for Picture Vocabulary, $\beta = .184$, 95% CI [.14, .23]; for Verbal Analogies, $\beta = .338$, 95% CI [.29, .39]) than at 225% of the poverty threshold (an income-to-needs ratio of 2.25; for Passage Comprehension, $\beta = .173$, 95% CI [.13, .21]; for Picture Vocabulary, $\beta = .155$, 95% CI [.12, .19]; for Verbal Analogies, $\beta = .306$, 95% CI [.26, .34]). For students from more middle-class families (an income-to-needs ratio of 4, approximately the sample mean) the magnitude of the effect was even weaker (for Passage Comprehension, $\beta = .133$, 95% CI [.09, .16]; for Picture Vocabulary, $\beta = .115$, 95% CI [.08, .15]; for verbal reasoning, $\beta = .259$, 95% CI [.22, .29]).

A similar pattern was found for math. The magnitude of the relation between the math discrepancy score and later math achievement became progressively weaker from the poverty threshold, to 225% of the poverty threshold, to 400% of the

poverty threshold ($\beta = .140$, 95% CI [.10, .19], $\beta = .107$, 95% CI [.07, .14], $\beta = .061$, 95% CI [.03, .09], respectively), holding all other variables constant.

Discussion

Using data from a 10-site, longitudinal study of U.S. children, the present study shows that students' academic achievements in high school are affected by early teacher expectations, such that high school students whose first-grade teachers underestimated their abilities performed significantly worse on standardized tests of math, reading comprehension, vocabulary knowledge and verbal reasoning than would have been predicted on the basis of their early test scores. Conversely, when early abilities were overestimated, high school students performed better than expected.

The findings of the present study demonstrate that misperceptions of abilities early in students' schooling continue to exert an effect on academic achievement 10 years later. This confirms the durability of self-fulfilling prophecies in the classroom reported by Alvidrez and Weinstein (1999), who found that preschool teacher expectations have an effect on high school GPA, though with a small sample size, a lack of information on noncognitive covariates, and an antiquated IQ test (Jussim et al., 2009). The present study, which used a large, nationally representative sample, included noncognitive controls collected with maternal reports and observational measures, and a robust measure of achievement and cognitive abilities, accounted for these limitations. The sizes of the effects found in the present study (as measured by the change in R^2 s and the standardized coefficients) are comparable to those found in other naturalistic studies on self-fulfilling prophecies in the classroom, where effects range from 0 to .40 and a weighted average of .07 (Jussim et al., 2009).

Table 6
Regression Analyses Predicting WJ-R Passage Comprehension, Picture Vocabulary, and Verbal Analogies Scores Using Language Skills Discrepancy Scores as Predictor

Predictor	WJ-R Passage Comprehension ^a			WJ-R Picture Vocabulary ^b			WJ-R Verbal Analogies ^c		
	B	SE B	β	B	SE B	β	B	SE B	β
Step 1	441.536	7.575		304.919	13.695		361.825	16.345	
WJ-R, 4.5 years	0.211	0.020	.355***	0.464	0.030	.501***	0.331	0.035	.331***
Child competence	-0.171	0.253	-.021	0.074	0.237	.009	0.786	0.283	.088**
Self-control	0.281	0.121	.075*	0.140	0.115	.036	0.422	0.137	.100**
Gender (female)	-0.022	0.795	-.001	-1.874	0.749	-.072*	0.403	0.894	.014
Ethnicity (African American)	-8.695	1.199	-.239***	-7.893	1.171	-.208***	-7.874	1.397	-.192***
Family income-to-needs ratio	0.486	0.136	.123***	0.254	0.128	.061*	0.527	0.152	.118***
Step 2	451.997	7.978		307.581	13.613		368.744	15.733	
WJ-R, 4.5 years	0.190	0.021	.320***	0.464	0.029	.500***	0.330	0.034	.330***
Child competence	-0.414	0.259	-.052	-0.146	0.244	-.018	0.212	0.282	.024
Self-control	0.251	0.120	.067*	0.110	0.114	.028	0.343	0.132	.081*
Gender (female)	-0.329	0.792	-.013	-2.187	0.749	-.084**	-0.412	0.865	-.015
Ethnicity (African American)	-8.753	1.188	-.240***	-7.819	1.162	-.206***	-7.681	1.343	-.188***
Family income-to-needs ratio	0.500	0.134	.126***	0.236	0.127	.057	0.480	0.147	.108*
Language skills discrepancy score	1.861	0.482	.131***	1.520	0.439	.103**	3.951	0.508	.247***
Step 3	453.193	7.963		309.353	13.585		370.732	15.704	
WJ-R, 4.5 years	0.188	0.021	.317***	0.461	0.029	.497***	0.327	0.034	.327***
Child competence	-0.454	0.258	-.057	-0.184	0.244	-.022	0.170	0.282	.019
Self-control	0.246	0.120	.066*	0.106	0.114	.027	0.339	0.132	.080*
Gender (female)	-0.300	0.789	-.012	-2.170	0.746	-.083**	-0.393	0.863	-.014
Ethnicity (African American)	-8.618	1.185	-.237***	-7.713	1.159	-.204***	-7.562	1.339	-.185***
Family income-to-needs ratio	0.511	0.134	.129***	0.246	0.126	.060*	0.492	0.146	.110**
Language skills discrepancy score	1.828	0.481	.129***	1.478	0.438	.100***	3.905	0.507	.244***
Language Skills Discrepancy Score \times Income	-0.393	0.157	-.078*	-0.366	0.149	-.069**	-0.410	0.172	-.072*

Note. WJ-R = Woodcock-Johnson—Revised Test of Achievement and Cognitive Abilities.

^a Step 1: $F(6, 717) = 49.87, p < .001, R^2 = .294$; Step 2: $F(7, 716) = 45.70, p < .001, \Delta R^2 = .014^{***}$; Step 3: $F(8, 715) = 41.06, p < .001, \Delta R^2 = .006^*$. ^b Step 1: $F(6, 718) = 86.78, p < .001, R^2 = .420$; Step 2: $F(7, 717) = 77.22, p < .001, \Delta R^2 = .010^{***}$; Step 3: $F(8, 716) = 68.80, p < .001, \Delta R^2 = .005^*$. ^c Step 1: $F(6, 720) = 49.35, p < .001, R^2 = .291$; Step 2: $F(7, 719) = 54.44, p < .001, \Delta R^2 = .055^{***}$; Step 3: $F(8, 718) = 48.66, p < .001, \Delta R^2 = .005^*$.

* $p < .05$. ** $p < .01$. *** $p < .001$.

The present study also suggests that teachers' under- and overestimation of students' early abilities disproportionately affects the high school math achievement of children from relatively poorer families. This result extends previous findings (a) that family income moderated the relation between sixth-grade teachers' perceptions of effort and students' final grades that year, as well as standardized math test scores in seventh grade (Jussim et al., 1996), and (b) that family income moderated the relation between first-grade teacher expectations and third-grade math achievement (Hinnant et al., 2009). Hinnant and colleagues, who used the same population and included covariates similar to those in the present study, reported a larger magnitude of self-fulfilling effects on math achievement for students from lower income families, (standardized coefficient = .20) than found in the present study (standardized coefficient = .14), consistent with research showing that self-fulfilling effects dissipate over time (A. E. Smith, et al., 1999; West & Anderson, 1976).

Moreover, this study evaluated the effect of teachers' under- and overestimation of language skills separate from basic reading abilities. For language skills, a significant interaction was found between the teacher discrepancy score and family income. Under- and overestimation of language skills disproportionately affected children from relatively poorer families with respect to reading comprehension, word knowledge, and verbal reasoning test scores.

This finding is consistent with the premise that vocabulary and phonological awareness are important for later reading and verbal reasoning abilities, particularly for students from lower income families (Hart & Risley, 1995; NICHD Early Child Care Research Network, 2005b; Stanovich, 1986, 2000). This differentially deleterious effect on poor students is not due to the impact of ethnicity on teacher expectations, because this variable was taken into account in the analyses. A teacher's under- and overestimation of a student's basic reading abilities did not differ by child income, gender, or ethnicity.

The present findings add insight to the relation between family income and children's academic achievement and abilities. Family income in early childhood is often a robust predictor of later achievement (Duncan & Brooks-Gunn, 1994; Entwisle, Alexander, & Olson, 2005). The magnitude of the effect of income on high school performance found in the present study, however, ranged from nonsignificant to .26 across the different levels of the teacher discrepancy scores and academic subjects. There is little research on factors that moderate the relation between family income and children's academic outcomes (Bradley & Corwyn, 2002). Based on converging literature on resiliency and on mediating mechanisms of SES on child outcomes, Bradley and Corwyn suggested that external support systems and access to additional resources are potential moderators of the association between SES

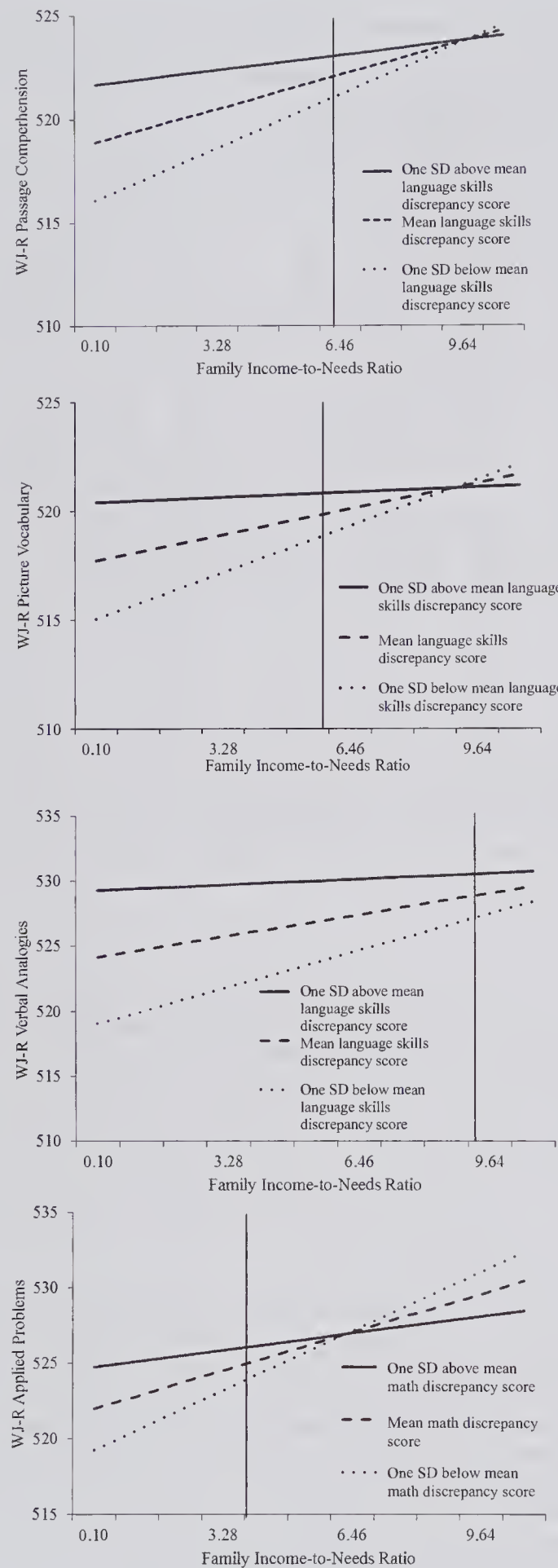


Figure 1 (opposite)

and children's academic outcomes, which are similar to possible mechanisms of self-fulfilling prophecies in the classroom (discussed below). This points to the importance of using integrated theoretical perspectives that consider the relations between mediation and moderation when investigating the developmental effects of SES (Baron & Kenny, 1986; Lerner, 2003).

Although the primary goal of the present study focused on the long-term effects of self-fulfilling prophecies on academic achievement, the findings of the preliminary analyses have implications for the study of biased teacher expectations. One issue is whether teachers' expectations are influenced by stereotypes. The present results suggest that teachers' perceptions of young students' math, basic reading, and language skills are mostly independent of students' demographic characteristics—with the exception that teachers tend to overestimate girls' language abilities. This is consistent with other research that has found that teachers' expectations are generally unrelated to students' demographic characteristics (Dusek & Joseph, 1983; Jussim et al., 1996; Madon et al., 1998) and that teachers tend to overestimate the amount of effort girls put forth in academics (Jussim et al., 1996; Madon et al., 1998). But many other studies have found that teachers' expectations can be biased based on a student's membership in stigmatized groups (Alvidrez & Weinstein, 1999; Jussim & Eccles, 1992; Madon et al., 1998). Some caution should be used when interpreting the results of the preliminary analyses because the present study did not consider teacher and classroom characteristics, which have been found to moderate the relation between student ethnicity and teachers' expectations (McKown & Weinstein, 2008; Rubie-Davies, 2006). Furthermore, the present study did not examine the accumulation of repeated beliefs or the relation of teacher expectations and stereotype threat, which may contribute to the association between teacher expectations and educational inequalities (Madon, Willard, Gyll, & Scherr, 2011; McKown, Gregory, & Weinstein, 2010). The results of the present study, however, suggest that teacher expectations are more influenced by a student's behavior in the classroom (i.e., self-reliance and positive affect) than by demographic characteristics.

The present analyses do not identify the mechanisms that link teacher expectations in the first grade to students' achievement in high school. Past research suggests that teachers treat students differently based on their expectations (Brophy, 1983; Brophy & Good, 1970, 1974; Harris & Rosenthal, 1985; Rosenthal, 1974). Differential treatment could then lead to differences in the amount of information that high- and low-expectation students are exposed to, which expands or limits the amount of information they can learn (Jussim et al., 2009; Rubie-Davies, 2007). Differential treatment could also influence a student's motivation and self-perceptions, in turn, affecting a student's school performance

Figure 1 (opposite). Graph of predicted WJ-R scores at age 15 as a function of first-grade family income-to-needs ratio and first-grade teacher discrepancy scores. Lines are plotted at one standard deviation above the mean, at the mean, and at one standard deviation below the mean of the teacher discrepancy scores. The vertical line represents the point on the income-to-needs ratio where the teacher discrepancy scores are no longer significant predictors of WJ-R scores. WJ-R = Woodcock-Johnson—Revised Test of Achievement and Cognitive Abilities.

(Brattesani et al., 1984; Jussim, 1989; Kuklinski & Weinstein, 2001).

To the best of my knowledge, this study is the first to consider teachers' misperceptions of young students' vocabulary and phonological awareness, so the specific mediating processes of self-fulfilling effects within this subject have not been explored. The direct effect of information exposure—or a better climate that somehow makes learning easier—at the beginning of an academic career may be particularly important, given that misperceptions of language skills were stronger predictors of age 15 WJ–R scores for students from lower income families than for more affluent students. Differential exposure has been linked to differences in language development between children from low- and high-income families (Hart & Risley, 1995; Hoff, 2003), and as children move through elementary school the gap between poor and proficient students tends to become larger (Stanovich, 1986). Perhaps early exposure to more advanced language information, relative to actual abilities, based on a teacher's misconception allows students from lower income families to catch up to their more affluent peers. Indeed, the quality of language exposure in early schooling is related to later language and reading development (Dickinson & Porche, 2011), exemplifying one way that improvements in language skills early in a student's school career could be diversely influential on advanced reading, vocabulary, and verbal reasoning development. This may also explain why relatively higher income students' verbal reasoning abilities, the most complex of the outcomes in the present study, were affected by first-grade teachers' misperceptions of language skills.

It is also possible that parents' expectations contribute to the different outcomes of teacher expectations by family income. That is, affluent students may be more likely than their less affluent peers to have parents whose own, presumably higher, expectations for their children's achievement countermand the deleterious effects of teacher underestimations. Studies using Eccles's parent socialization model demonstrated that parents with higher incomes and more education tended to have higher academic expectations for their children and that high expectations, in turn, predicted children's academic achievement through parents' behavior (Eccles & Davis-Kean, 2005).

One counter to the interpretations of the findings presented here is the possibility that first-grade teachers are better at predicting their students' later achievement than are standardized tests alone because teachers are sensitive to aspects of children's motivation or competence that predict later achievement but that are not captured by standardized tests. If this were the case, then the results of the present analyses simply demonstrate the natural unfolding of scholastic competence over time, rather than the impact of self-fulfilling prophecies. This is unlikely for several reasons. First, noncognitive covariates were included in the model in order to address this concern, and no evidence was found for this argument. Second, if there are noncognitive factors that influence WJ–R scores, the influence of these variables should have been taken into account by the use of pre-first-grade WJ–R scores as covariates in the analyses. It seems unlikely that first-grade teachers are able to pick up on something that would influence later performance on the WJ–R that would not also have influenced the students' earlier performance on the same test. Finally, because teachers' expectations influence later achievement among low- and middle-class students, but not students from higher income

families (except for verbal reasoning abilities, where relatively higher income students also benefited from teacher overestimation), any unidentified variable that accurately influences teachers' expectations would therefore have to operate differentially at different family income levels, which is also unlikely.

The present study has several limitations that must be acknowledged. Although many covariates were included in the analyses, the results cannot be interpreted as causal. Also, the present analyses only considered differences between African American and White students from lower income families. The self-fulfilling effects of teachers' misperceptions across these ethnic groups within students from middle- and high-income families should be explored in future research. Furthermore, the present results should not be generalized to student SES in general because only family income was considered. It is important to consider multiple indicators of SES separately when investigating the influence of SES on development and achievement (Duncan & Magnuson, 2003). Future research should consider the moderation of both parents' education and occupation on the association between teachers' misperceptions and students' achievement. Finally, the size of the effects found here, as measured by the standardized coefficients (the betas reported in the Results section) and the changes in R^2 , are small to moderate by traditional standards (Cohen et al., 2003).

The results of this study, however, should not be dismissed as practically unimportant on the basis of traditional measures of effect sizes, for several reasons. First, the changes in R^2 s when the interaction terms were entered into the equations in the present study were larger than the median effect size among over 100 articles that included interactions using multiple regression in influential applied psychology and business journals (Aguinis et al., 2005). Second, academic achievement is a multiple-determined factor, which makes above-moderate effects for one covariate rare (Ahadi & Diener, 1989). Finally, readers should remember that effects that are dismissed as small in psychology are often larger than effects in the medical fields that are taken very seriously by both medical researchers and the general public (Rosenthal, 1990).

Conclusion

This study investigated one aspect of the complex cognitive and behavioral processes underlying student–teacher relationships and found that early inaccurate teacher expectations were a lasting contributor to later academic performance. Furthermore, the findings suggest that self-fulfilling prophecies in the classroom vary across academic subjects and family income. Under- and overestimation of early math and language abilities, but not reading abilities, seemed to have a more meaningful effect on students from lower income families. The fact that self-fulfilling prophecies in first-grade classrooms exerted an especially lasting impact on the achievement of disadvantaged students raises the possibility that teachers' underestimation of poor children's academic abilities may be one factor that contributes to the persistent and worrisome gap in achievement between children from different socioeconomic backgrounds. On the other hand, teachers' overestimation of abilities seemed to disproportionately help low-income students, suggesting that knowledge of self-fulfilling prophecies in the classroom could be relevant to policies aimed at ameliorating the achievement gap between low- and high-income students,

especially considering the persistence of the achievement gap in America.

References

- Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94–107. doi:10.1037/0021-9010.90.1.94
- Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56*, 398–406. doi:10.1037/0022-3514.56.3.398
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- Alvidrez, J., & Weinstein, R. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*, 731–746. doi:10.1037/0022-0663.91.4.731
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182. doi:10.1037/0022-3514.51.6.1173
- Baumann, J. F., Kame'enui, E. J., & Ash, G. E. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, D. Lapp, J. R. Squire, & J. M. Jenson (Eds.), *Handbook of research on teaching the English language arts* (2nd ed., pp. 752–785). Mahwah, NJ: Erlbaum.
- Blachman, B. A. (2000). Phonological awareness. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 483–502). Mahwah, NJ: Erlbaum.
- Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*(1), 371–399. doi:10.1146/annurev.psych.53.100901.135233
- Brattesani, K. A., Weinstein, R. S., & Marshall, H. H. (1984). Student perceptions of differential teacher treatment as moderators of teacher expectation effects. *Journal of Educational Psychology, 76*, 236–247. doi:10.1037/0022-0663.76.2.236
- Brophy, J. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*, 631–661. doi:10.1037/0022-0663.75.5.631
- Brophy, J., & Good, T. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology, 61*, 365–374. doi:10.1037/h0029908
- Brophy, J., & Good, T. (1974). *Teacher–student relationships: Causes and consequences*. New York, NY: Holt, Rinehart & Winston.
- Burchinal, M., McCartney, K., Steinberg, L., Crosnoe, R., Friedman, S. L., McLoyd, V., . . . NICHD Early Child Care Research Network. (2011). Examining the Black–White achievement gap among low-income children using the NICHD Study of Early Child Care and Youth Development. *Child Development, 82*(5), 1404–1420. doi:10.1111/j.1467-8624.2011.01620.x
- Bus, A. G., & van IJzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 91*, 403–414. doi:10.1037/0022-0663.91.3.403
- Cohen, J., Cohen, P., West, C., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development, 82*(3), 870–886. doi:10.1111/j.1467-8624.2011.01576.x
- Duncan, G. J., & Brooks-Gunn, J. (1994). Economic deprivation and early childhood development. *Child Development, 65*, 296–318. doi:10.2307/1131385
- Duncan, G. J., & Magnuson, K. (2003). *Off with Hollingshead: Socioeconomic resources, parenting and child development*. Mahwah, NJ: Erlbaum.
- Dusek, J., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75*, 327–346. doi:10.1037/0022-0663.75.3.327
- Eccles, J., & Davis-Kean, P. (2005). Influences of parents' education on their children's educational attainments: The role of parent and child perceptions. *London Review of Education, 3*, 191–204. doi:10.1080/14748460500372309
- Eder, D. (1981). Ability grouping as a self-fulfilling prophecy: A micro-analysis of teacher–student interaction. *Sociology of Education, 54*(3), 151–162. doi:10.2307/2112327
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2005). First grade and educational attainment by age 22: A new story. *American Journal of Sociology, 110*(5), 1458–1502. doi:10.1086/428444
- Farkas, G., & Beron, K. (2004). The detailed age trajectory of oral vocabulary knowledge: Differences by class and race. *Social Science Research, 33*(3), 464–497. doi:10.1016/j.ssresearch.2003.08.001
- Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin, 97*, 363–386. doi:10.1037/0033-2909.97.3.363
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.
- Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods, 41*(3), 924–936. doi:10.3758/BRM.41.3.924
- Hecht, S. A., Burgess, S. R., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2000). Explaining social class differences in growth of reading skills from beginning kindergarten through fourth-grade: The role of phonological awareness, rate of access, and print knowledge. *Reading and Writing, 12*(1–2), 99–128. doi:10.1023/A:1008033824385
- Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*, 662–670. doi:10.1037/a0014306
- Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368–1378. doi:10.1111/1467-8624.00612
- Johnston, R. S., Anderson, M., & Holligan, C. (1996). Knowledge of the alphabet and explicit awareness of phonemes in pre-readers: The nature of the relationship. *Reading and Writing, 8*(3), 217–234. doi:10.1007/BF00420276
- Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology, 57*, 469–480. doi:10.1037/0022-3514.57.3.469
- Jussim, L., & Eccles, J. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology, 63*, 947–961. doi:10.1037/0022-3514.63.6.947
- Jussim, L., & Eccles, J. (1995). Naturally occurring interpersonal expectancies. In N. Eisenberg (Ed.), *Social development. Review of personality and social psychology* (Vol. 15, pp. 74–108). Thousand Oaks, CA: Sage.
- Jussim, L., Eccles, J., & Madon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 281–388). San Diego, CA: Academic Press. doi:10.1016/S0065-2601(08)60240-3
- Jussim, L., & Harber, K. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131–155. doi:10.1207/s15327957pspr0902_3
- Jussim, L., Robustelli, S., & Cain, T. (2009). Teacher expectations and

- self-fulfilling prophecies. In K. W. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 349–380). New York, NY: Routledge.
- Kuklinski, M. R., & Weinstein, R. S. (2001). Classroom and developmental differences in a path model of teacher expectancy effects. *Child Development*, 72, 1554–1578. doi:10.1111/1467-8624.00365
- Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.
- Lerner, R. M. (2003). What are SES effects effects of? A developmental systems perspective. In M. H. Bornstein & R. H. Bradley (Eds.), *Socio-economic status, parenting, and child development* (pp. 231–255). Mahwah, NJ: Erlbaum.
- Locke, A., Ginsborg, J., & Peers, I. (2002). Development and disadvantage: Implications for the early years and beyond. *International Journal of Language & Communication Disorders*, 37(1), 3–15. doi:10.1080/13682820110089911
- Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology*, 72, 791–809. doi:10.1037/0022-3514.72.4.791
- Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class, and ethnic stereotypes: A naturalistic study in person perception. *Personality and Social Psychology Bulletin*, 24(12), 1304–1318. doi:10.1177/01461672982412005
- Madon, S., Willard, J., Gyll, M., & Scherr, K. C. (2011). Self-fulfilling prophecies: Mechanisms, power, and links to social problems. *Social and Personality Psychology Compass*, 5(8), 578–590. doi:10.1111/j.1751-9004.2011.00375.x
- McGrew, K. S. (1993). The relationship between the Woodcock-Johnson Psycho-Educational Battery—Revised Gf-Gc cognitive clusters and reading achievement across the life-span. *Journal of Psychoeducational Assessment Monograph Series: WJ-R Monograph*, 39–53.
- McGrew, K. S., & Hessler, G. L. (1995). The relationship between the WJ-R Gf-Gc cognitive clusters and mathematics achievement across the life-span. *Journal of Psychoeducational Assessment*, 13, 21–38. doi:10.1177/073428299501300102
- McGrew, K. S., & Knopik, S. N. (1993). The relationship between the WJ-R Gf-Gc cognitive clusters and writing achievement across the life-span. *School Psychology Review*, 22, 687–695.
- McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *WJ-R technical manual*. Allen, TX: DLM.
- McKown, C., Gregory, A., & Weinstein, R. (2010). *Expectations, stereotypes, and self-fulfilling prophecies in classroom and school life*. Mahwah, NJ: Erlbaum.
- McKown, C., & Weinstein, R. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology*, 46(3), 235–261. doi:10.1016/j.jsp.2007.05.001
- Merton, R. (1948). The self-fulfilling prophecy. *The Antioch Review*, 8(2), 193–210. doi:10.2307/4609267
- NICHD Early Child Care Research Network. (2005a). *Child care and child development: Results from the NICHD Study of Early Child Care and Youth Development*. New York, NY: Guilford Press.
- NICHD Early Child Care Research Network. (2005b). Pathways to reading: The role of oral language in the transition to reading. *Developmental Psychology*, 41, 428–442. doi:10.1037/0012-1649.41.2.428
- Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, 76, 85–97. doi:10.1037/0022-0663.76.1.85
- Rist, R. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, 40, 411–451.
- Rosenthal, R. (1974). *On the social psychology of the self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms*. New York, NY: MSS Modular.
- Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, 45(6), 775–777. doi:10.1037/0003-066X.45.6.775
- Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, 3, 176–179. doi:10.1111/1467-8721
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York, NY: Holt, Rinehart & Winston.
- Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1(3), 377–415. doi:10.1017/S0140525X00075506
- Rubie-Davies, C. M. (2006). Teacher expectations and student self-perceptions: Exploring relationships. *Psychology in the Schools*, 43(5), 537–552. doi:10.1002/pits.20169
- Rubie-Davies, C. M. (2007). Classroom interactions: Exploring the practices of high and low expectation teachers. *British Journal of Educational Psychology*, 77(2), 289–306. doi:10.1348/000709906X101601
- Schachter, F. F. (1979). *Everyday mother talk to toddlers: Early intervention*. New York, NY: Academic Press.
- Smith, A. E., Jussim, L., & Eccles, J. (1999). Do self-fulfilling prophecies accumulate, dissipate, or remain stable over time? *Journal of Personality and Social Psychology*, 77, 548–565. doi:10.1037/0022-3514.77.3.548
- Smith, M. L. (1980). Teacher expectations. *Evaluation in Education*, 4, 53–55. doi:10.1016/0191-765X(80)90015-4
- Spitz, H. H. (1999). Beleaguered Pygmalion: A history of the controversy over claims that teacher expectancy raises intelligence. *Intelligence*, 27(3), 199–234. doi:10.1016/S0160-2896(99)00026-4
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the the acquisition of literacy. *Reading Research Quarterly*, 21(4), 360–407. doi:10.1598/RRQ.21.4.1
- Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford Press.
- Stuart, M., & Coltheart, M. (1988). Does reading develop in a sequence of stages? *Cognition*, 30(2), 139–181. doi:10.1016/0010-0277(88)90038-8
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Allyn & Bacon.
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., . . . Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology*, 33, 468–479. doi:10.1037/0012-1649.33.3.468
- Weinstein, R. (1979). Student perceptions of teacher interactions with male high and low achievers. *Journal of Educational Psychology*, 71, 421–431. doi:10.1037/0022-0663.71.4.421
- Weinstein, R. (2002). *Reaching higher: The power of expectations in schooling*. Cambridge, MA: Harvard University Press.
- West, C., & Anderson, T. (1976). The question of preponderant causation in teacher expectancy research. *Review of Educational Research*, 46(4), 613–630.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery—Revised*. Allen, TX: DLM.

Received August 2, 2011

Revision received October 10, 2012

Accepted December 26, 2012 ■

Sex-Specific Differential Prediction of College Admission Tests: A Meta-Analysis

Franziska T. Fischer, Johannes Schult, and Benedikt Hell
University of Konstanz

This is the first meta-analysis that investigates the differential prediction of undergraduate and graduate college admission tests for women and men. Findings on 130 independent samples representing 493,048 students are summarized. The underprediction of women's academic performance ($d = 0.14$) and the overprediction of men's academic performance ($d = -0.16$) are generalizable, albeit small. Transferred onto a 4-point grading scale, women earn college grades that are 0.24 points higher than those of men with the same admission test result. Combining admission tests with indicators of previous academic achievements, such as high school grades, reduces the amount of under- and overprediction. Moderator analysis reveals that the underprediction of women's academic performance by admission tests is a problem of the past and present. Predictor differences as well as criterion differences are not associated with over- and underprediction. Rather, undergraduate college admission tests show more underprediction of women's academic performance than graduate admission tests. These results point to differences between undergraduate and graduate students, the latter being more selected.

Keywords: differential prediction, test bias, gender, sex differences, meta-analysis

Supplemental materials: <http://dx.doi.org/10.1037/a0031956.supp>

Every year, millions of people take standardized admission tests in order to be accepted into a college or university. The significant influence of the tests on this key aspect of society has induced a vast amount of research regarding the predictive power of admission tests. The raw correlation between the SAT and first-year college grade point averages (GPA) is 0.35 and increases to 0.53 after correction for range restriction¹ (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008). The Graduate Management Admission Test (GMAT; Kuncel, Credé, & Thomas, 2007), the Graduate Record Examination (GRE; Kuncel, Wee, Serafin, & Hezlett, 2010), and subject-specific admission tests in German-speaking countries (Hell, Trapmann, & Schuler, 2007) show similar results. Although predictive validity is necessary for high stakes testing (see Standards for Educational and Psychological Testing; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), it is not sufficient for its fairness.

A review of existing literature supports the conclusion that professionally constructed tests are not systematically biased

against minority group members in the prediction of academic performance (Linn, 1973; Sackett, Borneman, & Connelly, 2008; Young & Kobrin, 2001). There is, however, evidence that achievement test scores underpredict women's academic performance (Holden, 1989; Young & Kobrin, 2001). In other words, females with the same test scores as males earn better college grades on average. Efforts to summarize the literature in this field are more than 10 years old and restricted with regard to content and method. The present study overcomes these limitations by providing an up-to-date meta-analysis. International research results respecting both undergraduate and graduate college admissions are considered and for the first time, group-specific residuals from large-scale studies are summarized with the help of meta-analytic techniques.

Test Fairness and Test Bias in Predicting Subgroups

Regarding the definition of *test fairness* and *test bias* there has been some disagreement in the past. Today, there is consensus that bias relates more to statistical approaches, whereas fairness is a more value-laden concept (Meade & Fetzer, 2009). In the present study, we focus on a bias, which can emerge in the prediction of a subgroup's criterion as defined by Cleary (1968):

A test is biased for members of a subgroup of the population if in the prediction of a criterion, for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup.

This article was published Online First March 18, 2013.

Franziska T. Fischer, Johannes Schult, and Benedikt Hell, Department of Psychology, University of Konstanz, Konstanz, Germany.

This research was supported by the Federal Ministry of Education and Research Grant 01FP0930 and the European Social Fund of the European Union.

We thank Sabrina Strohmeier for her help coding articles and Lea Ludwig for her comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Franziska T. Fischer, Department of Psychology, University of Konstanz, P.O. Box 27, D-78457 Konstanz, Germany. E-mail: Franziska.Fischer@ls.kv.bwl.de

¹ Analyzing only admitted and enrolled students underestimates the true correlation, since admitted students tend to have a narrower range of test scores than the applicant pool. This problem can be addressed by correcting the correlation for range restriction. The Pearson-Lawley multivariate correction can be applied for this purpose (e.g., Gulliksen, 1950).

In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. (p. 115)

This approach is endorsed by the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, 2003) and the Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999). The Standards conclude that “no bias exists if the regression equations relating the test and the criterion are indistinguishable for the groups in question” (American Educational Research Association et al., 1999, p. 79).

Although Cleary (1968) and the Standards (American Educational Research Association et al., 1999) used the term *test bias* “the term *differential prediction* much more accurately describes what is assessed by the regression-based procedure for evaluating the across groups equality of the relationship between the test and the criterion” (Meade & Fetzer, 2009, p. 740). The present study follows this suggestion and applies the term *differential prediction*.

Differences Between Differential Prediction and Differential Validity

It is important to distinguish between differential validity and differential prediction, because they are obviously related but not identical concepts (Young & Kobrin, 2001). *Differential validity* determines whether the correlations between test results and a criterion are equal across various groups. In contrast, *differential prediction* refers to group differences in regression equations or in standard errors of estimates. Consequently, “equal correlations do not necessarily imply equal standard errors of estimate, nor do they necessarily imply equal slopes or intercepts” (Linn, 1978, p. 511). A test may predict the criterion with the same accuracy for different subgroups but may still underpredict one of these groups.

In empirical test evaluations, the employment of differential validity studies is much more widespread than the employment of differential prediction studies. Differences in prediction, however, have a more direct bearing on considerations of selection (Linn, 1982). The underpredicted group is particularly worrisome, because group members with low scores on the test may not be admitted even though they would perform well at college or university (Huff, Koenig, Treptau, & Sireci, 1999).

How to Measure Differential Prediction

Analyzing Differences in Regression Equations

Gulliksen and Wilks (1950) recommended computing separate regression lines for each group and analyzing the components of these regression models sequentially in three steps: (a) compare the standard errors of estimate; (b) test the slope differences, assuming that the errors are equal; and (c) test the intercept differences, assuming that the errors and the slope differences are equal. Since then, this procedure has often been used without step one, testing for differences in the errors of estimate (e.g., Bridgeman & Wendler, 1991; Cleary, 1968). In order to predict academic performance with college admission tests, Cleary implemented this procedure in 1968. From that time on, most of the corresponding studies have referred to Cleary (1968) and have called this statis-

tical procedure the Cleary approach (e.g., Linn, 1973; Meade & Tonidandel, 2010). Performing a moderated multiple regression and testing its components is equivalent to this procedure (Bartlett, Bobko, Mosier, & Hannan, 1978). The corresponding formula is

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2, \quad (1)$$

in which X_1 is the predictor (e.g., admission test score), X_2 is the group dummy variable (e.g., sex), and \hat{Y} is the predicted criterion (e.g., predicted academic achievement). The group regression lines are regarded as being identical if there are no significant differences between the intercepts and/or slopes of the regression lines for the groups in question. If the intercepts differ but the slopes do not, one can draw a clear conclusion regarding over- and underprediction. The group with the larger intercept is underpredicted, meaning that members of this group perform better than predicted by the test (on average). The group with the smaller intercept is overpredicted by the test. These group members perform worse than predicted. If the slopes differ, regression lines may cross and therefore conclusions regarding over- and underprediction may vary between different test score sections.

Analyzing Differences in Group-Specific Residuals

As an alternative to the Cleary approach, Lawshe (1983) introduced a simplified procedure. In this method, the mean residuals for every group are calculated based on a common regression line. Negative group residuals indicate overprediction. Positive group residuals indicate underprediction. In some exceptions, the prediction error is calculated by subtracting the actual from the predicted criterion or the algebraic sign of the residuals is changed intentionally in order to align the sign to the meaning (e.g., Bridgeman, McCamley-Jenkins, & Ervin, 2000; Clark & Grandy, 1984; Talento-Miller, 2008). In large-scale studies, group-specific residuals are preferred to the significance tests of Cleary because statistical tests are prone to indicate significance due to the large test power (e.g., Cohen, 1988).

Previous Efforts to Summarize Sex-Specific Differential Prediction of Admission Tests

In this section, we provide a short summary of two previous reviews that explore differential prediction of admission tests by gender, and we show the restrictions of these reviews.

Sanber and Millman (1987) summarized differential prediction effects associated with standardized achievement tests. The authors aggregated 38 studies including 147 samples in an unpublished meta-analysis. Results showed a significant mean slope difference ($M = -0.80$; $SD = 1.87$) and no significant intercept difference ($M = 0.20$; $SD = 2.31$). The method used to compare and aggregate the b values is questionable because there is no statistical justification for a simple summary of slope differences (Becker & Wu, 2007). Since Sanber and Millman (1987) found slope differences, the aggregated results allowed no clear conclusion about over- and underprediction. Therefore, they provided a descriptive summary. Of the summarized samples 81% reported equal slopes for males and females. In 53% of these cases, the intercepts were higher for females than for males, indicating marginal underprediction of women and marginal overprediction of

men. Still, the results did not allow conclusions about college admission tests in particular, only about achievement tests in general.

In 2001, Young and Kobrin published an extensive summary about the literature on differential prediction in American college admission (Young & Kobrin, 2001). The review arrived at the conclusion that the majority of studies reported underprediction of females. More precisely, the mean underprediction of women was about 0.06 grade points (based on a 0–4 scale). One limitation of this study was that these results were inferred without applying meta-analytic techniques. Therefore, despite the broad scope of the review, conclusions about the generalizability of the results cannot be drawn. Further, studies using a combination of test scores and high school grades as predictors were not considered separately.

The Present Study

There are two major goals in the present study. The first is to examine the general extent of the potential underprediction of women's academic performance and the potential overprediction of men's academic performance by undergraduate and graduate admission tests. As a related question, we investigate whether the combination of high school GPA (HGPA) and undergraduate admission tests, or undergraduate GPA (UGPA) and graduate admission tests reduces the magnitude of differential prediction. Previous research suggests that combining grades and test scores yields less bias than test scores alone (e.g., Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008).

Unlike the two studies summarized in the previous section, we focus on undergraduate *and* graduate tests, and we separately investigate predictions based on tests and tests combined with grades. We implement an international perspective by searching established literature databases that list primary studies from all over the world, like Web of Science and PSYINDEX. We also include recent findings, since many large-scale studies have been published in the last decade. Last, but not least, we apply meta-analytic methods to test the generalizability of the results.

If we find differential prediction, the second goal is to improve our understanding regarding the factors that are related to the underprediction of women's performance by identifying potential moderators. Various moderators are discussed in the literature (e.g., Sackett et al., 2008; Zwick, 2002). In the present meta-analysis we focus on the most promising.

First, we look at publication and sample properties. Nowadays items are well reviewed to avoid test content that is more familiar to men or women (Zwick, 2002, p. 152; Educational Testing Service, 2009). To investigate whether this fact has reduced differential prediction changes over time are examined. The mean age of the prospective students is assessed to control for possible gender differences in cognitive development (e.g., Ellis et al., 2008, p. 287; Lynn & Irwing, 2004; Lynn & Kanazawa, 2011).

Second, we look at test and grading properties. The test type/name is an obvious candidate regarding moderators because test content plays a crucial role in test outcomes (Zwick, 2002). Similarly, different test components such as verbal and mathematic sections might be linked to differential prediction (e.g., Bridgeman, Pollack, & Burton, 2008; Patterson, Mattern, & Kobrin, 2009).

We also test if differential prediction is related to test score differences between men and women and differences in college grades. According to Meade and Fetzter (2009), if a biased test is responsible for different group regression lines, there is most likely a test score difference (but no relating criterion difference). Conversely, criterion differences (but no relating test score difference) indicate some type of bias in the criterion.

There is also the possibility that the prediction of final grades is less biased than the prediction of the first-year GPA. This could be due to differential dropout rates or due to changing requirements of the learning content. We therefore also test the relation of differential prediction and the time span between predictor and criterion assessment.

Finally, we look at course taking patterns. It was argued that females tend to enroll in less stringent courses with more lenient grading systems (Alon & Gelbgiser, 2011; Conger & Long, 2010). Correcting for differences in grading standards or course taking patterns reduced underprediction of women (Bridgeman et al., 2000; Ramist, Lewis, & McCamley-Jenkins, 1994; Willingham, Pollack, & Lewis, 2002). Leonard and Jiang (1999), nevertheless, showed that underprediction of women's grades persists after controlling for gender differences in fields of study and for sample selection bias.

Method

Literature Search

We used three search strategies to locate published and unpublished studies: (a) database searches of PsycINFO, ERIC, PubMed, PsycARTICLES, Web of Science, PSYINDEX, and Google Scholar using the search terms (sex or gender) paired with (differential predict* or academic predict* or predict* bias etc.) paired with (admission test* or placement test* etc.); (b) manual searches through the reference lists of key articles; and (c) screenings of test-homepages and homepages of test providers (e.g., The College Board). The search was conducted at the beginning of 2010. All studies published before then were considered.

Inclusion Criteria

Each of the potential articles was evaluated for inclusion based on the following criteria. First, the study had to examine the prediction of men's and women's college performance by an admission test or a combination of admission test result and previous grades. Second, the authors had to report differential prediction results for men and women by (a) estimating separate regression lines for each gender and comparing their slopes and intercepts (this also includes moderated multiple regression studies with interaction terms); or (b) estimating a joint regression line, analyzing the mean residual for each gender and reporting enough information to calculate effect sizes; or (c) providing all required information to calculate the standardized mean residuals post hoc. Third, the study must have been published in English or in German. We also considered studies written in German, our native language, to further extend the number of potential samples. If the same sample was analyzed in multiple studies, we only included the study that contained most of the relevant data to avoid a duplicate study effect (Wood, 2008).

Summary of the Data Set

The literature search identified 962 studies. Out of this pool, 42 studies met all of the inclusion criteria. The remaining studies could not be included, mainly because they only reported differential validities without providing statistics on differential prediction. Further reasons for exclusion were limited criteria information (e.g., dichotomous pass/fail) and insufficient information about the required statistics for each gender. Also, prediction models that contained additional predictor variables (e.g., personality traits) could not be statistically disentangled.

The selected studies were published between 1973 and 2009, and they contained 130 samples with a total of 493,048 participants. Group-specific residuals or the information required to calculate them were reported in 28 studies (83 samples). Out of these samples, 55 reported residuals based on an admission test, and 52 offered residuals based on a combination of admission test and H GPA/UGPA. Differences in regression equations were reported in 14 studies (47 samples). Apparently, there was an overlap between the studies/samples reporting residuals based on admission tests and studies reporting residuals based on the combination of test scores and H GPA/UGPA. We handled this dependency by separately aggregating the residuals. There was no overlap between samples providing residuals and samples providing differences in regression equations. The criterion was typically first year GPA (more detailed characteristics of the studies, such as author, sample size, and name of the admission test are presented in Tables S1 and S2 in the online supplemental materials).

Coding of Study Variables

Data of independent samples were coded separately. For some samples, all required information was obtained for different predictors and/or criteria. In these cases, the following decision rules were applied. First, the whole test was used as the predictor instead of test parts. Second, the criterion with the biggest sample size was chosen. In ambiguous cases, the first year GPA was analyzed instead of later earned grades. Only three studies offered different criteria. We therefore could not make use of multivariate techniques to handle multiple outcomes (e.g., Becker, 2000).

Following aspects were coded as potential moderators: publication year, age of the participants, test type, verbal and mathematic test components, gender differences in test scores, gender differences in H GPA/UGPA, average time span between predictor assessment and criterion assessment, and freedom of course choice. Test score and grade differences were expressed in effect sizes before the moderator analyses were performed. If both statistics were given, we corrected the predictor differences with the help of the criterion differences and vice versa.

The first and the second authors coded all of the studies independently. Both were familiar with the field of study and had created the coding scheme. The initial interrater agreement was 96%. Discrepancies between the raters were solved by consulting a third rater and having discussions to reach a consensus. There were no coded variables with a disproportionate amount of initial differences.

Analytical Procedures

Summarizing residuals. In order to aggregate residuals, they have to be transferred to summable statistics. Lawshe (1983) proposed to test whether group-specific mean residuals (\bar{E}_{men} and \bar{E}_{women}) differ with

$$t = [(\bar{E}_{men} - \bar{E}_{women})/SD] \cdot \sqrt{N}. \quad (2)$$

Unfortunately, the two mean residuals are not independent of each other because

$$N_{men} \cdot \bar{E}_{men} + N_{women} \cdot \bar{E}_{women} = 0. \quad (3)$$

Thus, the assumption of the t test for independent subgroups is violated. To avoid this problem, we do not agree with the proposal from Lawshe (1983). Instead, we recommend the null hypothesis that suggests that the sex-specific residuals do not differ from zero:

$$t_j = [(\bar{E}_j - 0)/SD] \cdot \sqrt{N_j} \quad (j = men, women). \quad (4)$$

Previous research suggests an underprediction of women's performance, but we do not want to exclude the option that there is indeed an overprediction. The same applies vice versa to men. Therefore, we recommend looking at two-tailed tests.

In the present meta-analysis we did not perform the t tests but, rather, calculated the corresponding effect sizes within each sample (for each gender separately). According to Cohen (1988, pp. 45–48) the effect sizes were calculated by

$$d_j = (\bar{E}_j - 0)/SD \quad (j = men, women). \quad (5)$$

We used the standard deviation of the total sample since the residuals are based on one regression line (this procedure is equivalent to standardizing the residuals). If the total standard deviation was not reported, we computed the pooled standard deviation. In cases where there was no standard deviation available, it was calculated by

$$SD = \sqrt{S_y^2 \cdot (1 - R_{xy}^2)}, \quad (6)$$

in which S_y^2 is the criterion variance and R_{xy}^2 is the variance explained by the regression.

After calculating effect sizes within each sample, we separately accumulated the d values for men and women. Two bare-bones meta-analyses were performed (i.e., correction of the observed variance for sampling error; Hunter & Schmidt, 2004). We chose the random-effects model as we expected effect size variations based on sample characteristics (Borenstein, Hedges, Higgins, & Rothstein, 2010). Significant Q statistics in the fixed-effects tests of homogeneity should support this choice. The corresponding formula for the weighted average effect size was

$$\bar{d}_j = \sum w_{ji} d_{ji} / \sum w_{ji} \quad (j = men, women; i = sample), \quad (7)$$

with w_i as the weight for the i th study. The inverse of the variance, which for one variable relationship is sample size divided by the variance of the target variable, was used as the weight (Lipsey & Wilson, 2001, p. 72). Since we had standardized effect sizes d (and not raw mean residuals) the standard deviation of the target vari-

able was 1, so $w_i = n_i$. For a corresponding example, see Hunter and Schmidt (2004, p. 289). Further corrections for artifacts were not possible because there were no artifact information reported with regard to the mean residuals.

We calculated 95% confidence intervals and 90% credibility intervals and tested the homogeneity of the effect sizes. The homogeneity test allows conclusions on whether the samples do share a common population effect size or not. We used the heterogeneity test (Q test) according to Shadish and Haddock (2009) based on a fixed-effects model. Finally, we conducted moderator analyses to examine the source of heterogeneity.

Summarizing regression equations. Although the methodological literature on meta-analytic techniques is substantial, little attention has been paid to the issue of summarizing regression slopes and intercepts. This fact can be explained by several challenges (for a detailed discussion, see Aguinis, Culpepper, & Pierce, 2010).

An overview of the existing methods for summarizing slopes was given by Becker and Wu (2007). They addressed the shortcomings of these methods by presenting a new multivariate generalized least squares approach. Anyhow, this method is also limited, because it requires knowledge of the covariances among slopes, which are rarely reported.

A new approach recommended using the semipartial correlation (between predictor and criterion) as a partial effect size (Aloe & Becker, 2011; e.g., see Aloe & Becker, 2009). This method allows summarizing linear models with more than one predictor. In the present model, we have two predictors and an interaction term (see Equation 1). To aggregate the interaction terms the corresponding t -statistic as well as the correlation between the interaction term and the test score is needed (given that, the test score and sex are related). Our identified studies rarely reported this information (especially the correlations). Moreover, some studies only reported the standardized regression weight for gender (e.g., Bridgeman & Wendler, 1991) and others reported the contribution to R^2 of intercept and slopes (e.g., Pennock-Román, 1994). This mixture of available information about regression equations is in line with Borneman's (2010) conclusion that "it is unlikely there are sufficient data in published manuscripts lying around for meta-analysis" (p. 225). Despite theoretically having the desired statistical properties, methods for aggregating regression equations could not be applied because the relevant studies did not report sufficient data. In order to still summarize the studies reporting regression equations, we created a descriptive summary.

Results

Gender-Specific Residuals

For each gender, we calculated separate mean effect sizes based on (a) studies that used admission test results as the sole predictor and (b) studies that used a combination of admission test results and HGPA or UGPA as the predictor. We also calculated mean effect sizes across the studies, without the four large-scale studies ($N > 5,000$). The results were substantially the same.

Egger's regression test for funnel plot asymmetry (Sterne & Egger, 2005) is not significant for three of the four funnels (females: admission test as predictor $t = 1.23, p = .225$; females: admission test and HGPA/UGPA as predictor $t = 1.90, p = .063$; males: admission test and HGPA/UGPA as predictor $t = 0.93, p = .356$). Only the funnel of the effect sizes for males, based on the admission test as the sole predictor, reaches significance ($t = 2.52, p = .015$). Since the effect sizes for males and females are based on the exact same amount of unpublished and published studies, we think there is no substantial publication bias. Moreover, the asymmetric funnel is related to an unequal amount of men and women in some samples. There are two outliers in the funnel identified as asymmetric. Both samples show a very low percentage of men (less than 28%). Following Equation 3, residuals are not independent of sample sizes. Particularly a small proportion of one group within one sample can result in an extreme mean residual for this group. This is the case with these two outliers.

Admission test as predictor. Table 1 shows the corresponding mean effect sizes for women ($d_{female} = 0.14$, indicating underprediction) and men ($d_{male} = -0.16$, indicating overprediction). Because neither confidence nor credibility intervals overlap zero, the results can be generalized (Schmidt & Hunter, 1982). According to a fixed effect model the effect sizes are heterogeneous for women, $Q_{fix}(54) = 111.31, p < .001$, and homogenous for men $Q_{fix}(54) = 69.49, p = .076$.

Admission test combined with HGPA or UGPA as predictor. For studies using admission tests and HGPA/UGPA as predictors, mean effect sizes are slightly smaller ($d_{female} = 0.11$ and $d_{male} = -0.12$; see Table 1). It must be kept in mind that these results are not completely independent from the analyses of the studies that were using only admission tests as the predictor due to overlapping samples. Again, the confidence intervals as well as the credibility intervals do not include zero. The heterogeneity analyses reveal that the effect sizes are heterogeneous for women, $Q_{fix}(50) = 76.75, p < .01$, and homogeneous for men, $Q_{fix}(51) = 49.76, p = .523$.

Table 1
Differential Prediction Effects for Women and Men

Predictor	Women					Men				
	<i>k</i>	<i>N</i>	<i>d</i>	95% CI	90% CRI	<i>k</i>	<i>N</i>	<i>d</i>	95% CI	90% CRI
Admission test	55	154,162	0.14	[0.13, 0.16]	[0.08, 0.21]	55	140,950	-0.16	[-0.17, -0.15]	[-0.20, -0.13]
Admission test and HGPA/UGPA	51	220,321	0.11	[0.10, 0.12]	[0.07, 0.14]	52	203,940	-0.12	[-0.12, -0.11]	[-0.17, -0.06]

Note. *k* = number of samples; CI = confidence interval; CRI = credibility interval; HGPA = high school grade point average; UGPA = undergraduate grade point average. Positive effect sizes indicate underprediction, whereas negative effect sizes indicate overprediction.

Moderator analysis. Effect sizes are heterogeneous for women based on admission tests as the sole predictor. To explain this effect size distribution for women we conducted analyses for potential moderators.

Test type is a significant moderator, $Q_{between}(5) = 76.97, p < .001$. The effect sizes for the different tests where more than one sample was available are displayed in Table 2. The underprediction of women's academic performance is negligible for the graduate admission tests, GRE and Medical College Admission Test (MCAT), and small to moderate for the undergraduate admission tests, SAT² ($d_{female} = 0.14$) and ACT ($d_{female} = 0.30$).

Separate prediction results for mathematics and verbal test sections were reported by four large-scale SAT studies. The mathematics section shows more underprediction of women ($d_{female} = 0.17, k = 4, N_{female} = 135,144, 95\%$ confidence interval [CI] = [0.15, 0.19], 90% credibility interval [CRI] = [0.15, 0.19]) than the verbal section ($d_{female} = 0.10, k = 4, N_{female} = 135,144, 95\%$ CI [0.09, 0.11], 90% CRI [0.10, 0.11]).

As shown in the previous analyses about the moderator test type, the study about the ACT (American College Testing Program, 1973) provides extreme over- and underprediction. At the same time, this study is by far the oldest one and the time span between predictor and criterion assessments is very short. We therefore tested the moderator variables, publication year and time span, with and without the ACT study. The results indicate the great influence of the ACT study. The significant influence of the moderators disappears if the ACT study is removed from the analyses. The influence of the remaining moderator variables predictor differences and criterion differences is not significant. Statistics for the moderator analyses are presented in detail in Table 3.

For the moderator variables, age and course choice, there were not enough data from the primary studies for the analysis.

Differences in Group Regression Equations

As described in the section *summarizing regression equations*, we present a descriptive summary of the studies offering group regression equations. Out of the included samples using admission tests as the sole predictor ($k = 20, N = 31,798$), 14 (70%) show significant slope and/or intercept differences, which indicate differential prediction. Eight of the samples showing differential prediction underpredict women's performance and overpredict men's performance. One sample shows no clear direction of the effect. The other five samples neither report conclusions about over-/underprediction nor report the required statistics to derive the information.

Predictions using a combination of admission test and H GPA or U GPA ($k = 35, N = 51,436$) show differential prediction less often. In 16 of these samples (46%), significant slope and/or intercept differences appear. Out of these samples, six underpredict women's performance, whereas one underpredicts men's performance. Unfortunately, the other nine samples do not report conclusions about overprediction and underprediction or the required statistics to derive the information.

Noticeably, the average sample size of studies reporting significant slope or intercept differences is higher ($N_{mean} = 2,032$) than the average sample size of the studies reporting no differences

($N_{mean} = 573$). This is not a surprise since significance depends, besides other factors, on sample size.

Discussion

The analysis of residuals shows that undergraduate and graduate admission tests underpredict women's academic performance ($d = 0.14$) and overpredict men's academic performance ($d = -0.16$), on average. According to Cohen's (1988) classification, these effect sizes are less than small. This classification was an initial general attempt and not intended to be applied to every situation. Less than small underprediction may still have tangible consequences for admission decisions. Aguinis, Beaty, Boik, and Pierce (2005) showed that this occurs frequently in studies of differential prediction.

When the effect sizes are transferred onto a 4-point grading scale (plugging in the mean standard deviation of residuals of the studies with the largest sample sizes), the academic performance of women is 0.11 points better than that predicted by the test. At the same time, men achieve grades that are 0.13 points worse than that predicted. In other words, with the same admission test result, women earn 0.24 points better grades than men do. The amount of underprediction and overprediction is smaller when admission tests are used in combination with H GPA/UGPA ($d_{female} = 0.11, d_{male} = -0.12$).³ In fact, the academic performance of women is 0.08 points better and the academic performance of men is 0.09 points worse than predicted. Taken together, our research confirms the findings of Young and Kobrin (2001), who report a mean underprediction of women's performance of 0.06 grade points. However, our results also show that the differential prediction effect is almost twice as big if the admission test is used as the sole predictor.

Studies comparing regression equations yield similar results. Samples in which admission tests are used as the sole predictor show differential prediction more often than those with a combination of admission test results and H GPA/UGPA (70% vs. 46%). The prevalent direction of the effect is underprediction of women's academic performance. The number of studies that find no differential prediction is surprisingly small when compared to the number of studies that show group-specific residuals around zero. This might be because of publication bias, that is, the tendency for null results to remain unpublished. Further, almost all samples used undergraduate admission tests as a predictor, whereas the studies that show group-specific residuals around zero are mostly based on graduate admission tests.

² The analysis includes older SAT versions as well as the revised SAT version, which includes a writing component.

³ This fact raises the question, whether H GPA or U GPA are biased in the opposite direction, that is, overpredicting women's academic performance. We analyzed the mean effect size of differential prediction for H GPA or U GPA for the included samples. The results show very small underprediction for women ($d_{female} = 0.07, k = 24, N_{female} = 144,383, 95\%$ CI [0.06, 0.09], 90% CRI [0.03, 0.12], $Q(23) = 50.99, p < .001$) and very small overprediction for men ($d_{male} = -0.08, k = 24, N_{male} = 131,675, 95\%$ CI [-0.09, -0.06], 90% CRI [-0.11, -0.04], $Q(23) = 38.93, p < .05$). In short, H GPA or U GPA seems to be biased in the same direction as admission tests, but the magnitude is attenuated.

Table 2
Differential Prediction Effects for Women Moderated by Test Name

Test name	Studies	<i>k</i>	<i>N_f</i>	<i>d_f</i>	95% CI	90% CRI	<i>Q_{within}</i>	<i>p</i>
SAT	6	7	139,856	0.14	[0.13, 0.15]	[0.12, 0.16]	5.67	.461
ACT	1	19	8,928	0.30	[0.25, 0.34]	[0.23, 0.36]	21.85	.239
GRE	5	13	2,589	0.03	[-0.02, 0.08]	[-0.11, 0.18]	4.85	.963
MCAT	1	14	1,312	0.02	[-0.02, 0.06]	[-0.11, 0.15]	1.96	.999

Note. Studies = number of studies included; *k* = number of samples; *d_f* = effect size for females based on admission test as predictor, whereas positive effect sizes indicate underprediction; CI = confidence interval; CRI = credibility interval; ACT = American College Test; GRE = Graduate Record Examination; MCAT = Medical College Admission Test.

Possible Reasons for the Underprediction of Women's Academic Performance

First, underprediction of women's performance remains an ongoing topic, as the differential prediction could not be reduced during the last decades though items are well reviewed for content fairness. The underprediction of women is further associated neither with test score differences (possibly indicating a bias in the test) nor with grade differences (possibly indicating some type of bias in the criterion; Meade & Fetzer, 2009). Consequently, disposing test score differences, by restructuring a test, does not necessarily reduce underprediction.

Different levels of over- and underprediction are rather related to different admission tests. Graduate admission tests indicate less of a problem with underprediction than undergraduate admission tests. This conclusion is consistent with the findings of Kuncel and Hezlett (2007). The underprediction linked to undergraduate tests might be explained by differences in *studying habits*. Women typically devote more effort to their academic work and show higher class attendance and greater academic motivation (Zwick, 2002). When accounting for such variables sex-specific differential prediction was reduced (Stricker, Rock, & Burton, 1993). As graduate students are a more *selective sample*, they possibly show

more homogeneous personality characteristics and skills across the sexes than high school alumni.

Another explanation might be more course choice possibilities during undergraduate studies. Unfortunately, there is not enough data to test the influence of course choice on differential prediction within the present study.

A further prominent explanation for the underprediction of women's academic performance is the influence of *stereotype threat* during the test execution. This means, women are under additional pressure that interferes with their test performance, because men are expected to outperform them on tests (e.g., Spencer, Steele, & Quinn, 1999; Steele, 1997). Still, the examination of stereotype threat in real world settings is difficult and is just at the beginning (Sackett, 2003; Sackett, Hardison, & Cullen, 2005). Given that we find differential prediction only in some admission tests makes it implausible that differential prediction is strongly associated with stereotype threat.

We also found differences between test components. The SAT verbal section shows less underprediction than the mathematics section. One explanation could be the *multiple-choice format*, which is more prevalent for mathematical than for verbal sections. Men tend to perform better in multiple-choice formats than women, whereas women reach at least equal scores in free-response formats (Bridgeman & Lewis, 1994; Lindberg, Hyde, Petersen, & Linn, 2010). Lack of time might be responsible for these differences (Goldstein, Haldane, & Mitchell, 1990).

Summing up, our results indicate that the underprediction of females' academic performance is not related to test score differences or criterion differences. Moreover, especially undergraduate admission tests are prone to differential prediction. Sex differences of undergraduate students with regard to their study habits and motivational factors are promising explanations that call for future investigations.

Strengths and Weaknesses of Methods Measuring Differential Prediction

Testing for differences in regression lines. Analysis of differences in regression lines is easy to illustrate and has been used for years. However, most of the relevant studies failed to report the information required to aggregate the results with meta-analytic techniques. Another pitfall concerns the intersection of regression lines. If regression lines intersect at a low predictor level, the intercept test can reveal underprediction of women, while the test score range containing most applicants overpredicts women (Schmidt & Hunter, 1982). To avoid this problem it is recommended to center the predictor variable or to define the areas

Table 3
Influence of Moderators on Differential Prediction Effects for Women

Moderator	<i>k</i>	β	<i>R</i> ²	<i>p</i>
Publication year	55	-.658	.43	<.001
Publication year ^a	36	.212	.04	.200
Predictor differences ^b	14	-.029	.00	.936
Criterion differences ^b	32	-.100	.01	.694
Time	44	-.344	.12	<.05
Time ^a	25	.314	.10	.085

Note. *k* = number of samples; time = time between admission test and criterion measure. *R*² = explained variance calculated conventionally following Lipsey and Wilson (2001). Studies that report insufficient data to code a particular moderator are omitted from that analysis; therefore, *k* fluctuates between analyses. Predictor and criterion differences are based on effect sizes, subtracting women's scores from the men's scores, respectively. Positive betas denote increases in women's effect size as the value of the predictor increases, whereas negative betas denote decreases in effect size as the value of the predictor increases.

^a Analysis without the American College Test study (American College Testing Program, 1973). ^b Predictor differences were corrected for criterion differences and vice versa, if the required statistics were given. We also performed the analyses without the corrections; the results were essentially the same.

where the group-specific regression lines differ. Only few studies implemented these recommendations. Therefore, it is possible that the results are artifacts and do not allow conclusions about the actual research questions.

Finally, finding significant differences between intercepts or slopes depends on sample size. In contrast to the meta-analytic approach, a descriptive overview cannot take this problem into account.

Reporting group-specific residuals. Reporting residuals helps to communicate test properties to a lay audience. Unstandardized mean residuals can be easily interpreted as the average deviation from the common regression line in the unit of the criterion scale. The residual method can be applied to large-scale studies and residuals can be transformed into effect sizes, which can be aggregated in meta-analysis. Despite these advantages, the method has its limitations.

The mean residual for women is not independent from the male residual and their sample sizes, respectively (see Equation 3). As a consequence, minorities reach more extreme mean residuals than the corresponding majority. When a common regression line is inappropriate, the analysis of residuals can be misleading as well. This is the case when slopes of group-specific regression lines have different algebraic signs, so that the lines intersect near the mean of the predictor (Norborg, 1984). In other words, one-half of each group (i.e., the upper or lower half on the test score scale) is overpredicted, whereas the other half is underpredicted. The co-existence of overprediction and underprediction remains undetected because both group mean residuals are zero.

Methodological conclusions. With regard to future meta-analysis of differential prediction results, we recommend that all relevant information in primary studies should be reported. Additionally, the scatter plots of the group-specific regression lines should be inspected to determine the curve progressions, especially potential intersections. In some cases, it might be helpful to further provide residual results for different predictor regions, for example, around the *cutoff point* used for admission. Additionally, the variance-covariance matrix should be provided to enable the meta-analysis of regression slopes (Borneman, 2010) as well as correlations (Aloe & Becker, 2011).

Final Conclusion

The present meta-analysis shows that admission tests underpredict women's academic performance and overpredict men's academic performance to a small but consistent extent. This conclusion holds true for older as well as for newer tests and is not related to predictor or criterion differences. Particularly undergraduate admission tests are more prone to over- and underprediction effects than graduate tests. Future research should build on these results. We suggest to focus on sex differences in noncognitive factors like study habits and motivational factors of undergraduate students rather than on test or criterion differences.

References

- References marked with an asterisk indicate studies included in the meta-analysis.
- Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94–107. doi:10.1037/0021-9010.90.1.94
- Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*, 648–680. doi:10.1037/a0018714
- Aloe, A. M., & Becker, B. J. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher, 38*, 612–624. doi:10.3102/0013189X09353939
- Aloe, A. M., & Becker, B. J. (2011). Advances in combining regression results in meta-analysis. In M. Williams & W. P. Vogt (Eds.), *The SAGE handbook of innovation in social research methods* (pp. 331–352). London, England: Sage.
- Alon, S., & Gelbgiser, D. (2011). The female advantage in college academic achievements and horizontal sex segregation. *Social Science Research, 40*, 107–119. doi:10.1016/j.ssresearch.2010.06.007
- *American College Testing Program. (1973). *Assessing students on the way to college: Vol. 1. Technical report for the ACT assessment program*. Iowa City, IA: Author.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31*, 233–241. doi:10.1111/j.1744-6570.1978.tb00442.x
- Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley, & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). San Diego, CA: Academic Press. doi:10.1016/B978-012691360-6/50018-5
- Becker, B. J., & Wu, M. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science, 22*, 414–429. doi:10.1214/07-STS243
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97–111. doi:10.1002/jrsm.12
- Borneman, M. J. (2010). Using meta-analysis to increase power in differential prediction analyses. *Industrial and Organizational Psychology, 3*, 224–227. doi:10.1111/j.1754-9434.2010.01228.x
- Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37–50. doi:10.1111/j.1745-3984.1994.tb00433.x
- *Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning test* (Research Report No. 2000–1). New York, NY: College Board.
- *Bridgeman, B., Pollack, J., & Burton, N. (2008). *Predicting grades in different types of college courses* (Research Report No. 2008–1). New York, NY: College Board.
- *Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology, 83*, 275–284. doi:10.1037/0022-0663.83.2.275
- *Burton, N. W., & Wang, M. (2005). *Predicting long-term success in graduate school: A collaborative validity study* (GRE Board Research Report No. 99-14R). Princeton, NJ: Educational Testing Service.
- *Calkins, D. S., & Whitworth, R. (1974). *Differential prediction of freshmen grade point average for sex and two ethnic classifications at a southwestern university*. Retrieved from ERIC database. (ED102199)
- *Casserly, P. L. (1982). *Older students and the SAT* (College Board Report No. 82–2). New York, NY: College Board.

- *Chou, T., & Huberty, C. J. (1990). *A freshman admissions prediction equation: An evaluation and recommendation*. Retrieved from ERIC database. (ED333081)
- *Clark, M. J., & Grandy, J. (1984). *Sex differences in the academic performance of Scholastic Aptitude Test takers* (College Board Report No. 84-8). New York, NY: College Board.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5, 115-124. doi:10.1111/j.1745-3984.1968.tb00613.x
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Conger, D., & Long, M. C. (2010). Why are men falling behind? Gender gaps in college performance and persistence. *Annals of the American Academy of Political and Social Science*, 627, 184-214. doi:10.1177/0002716209348751
- *Cowen, S., & Fiori, S. J. (1991, November). *Appropriateness of the SAT in selecting students for admission to California State University, Hayward*. Paper presented at the annual meeting of the California Educational Research Association, San Diego, CA. Retrieved from ERIC database. (ED343934)
- *Crawford, P. L., Alferink, D. M., & Spencer, J. L. (1986). *Postdictions of college GPAs from ACT composite scores and high school GPAs: Comparisons by race and gender*. Retrieved from ERIC database. (ED 326541)
- *Dlugosch, S. (2005). *Prognose von Studienerfolg dargestellt am Beispiel des Auswahlverfahrens der Bucerius Law School* [Prognosis of successful studies described by means of the selection method of the Bucerius Law School]. Aachen, Germany: Shaker.
- Educational Testing Service. (2009). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Author.
- *Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement*, 25, 333-347. doi:10.1111/j.1745-3984.1988.tb00312.x
- Ellis, L., Karadi, K., Hershberger, S., Field, E., Wersinger, S., Pellis, S., . . . Hetsroni, A. (2008). *Sex differences: Summarizing more than a century of scientific research*. New York, NY: Psychology Press.
- Goldstein, D., Haldane, D., & Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition*, 18, 546-550. doi:10.3758/BF03198487
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley. doi:10.1037/13240-000
- Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika*, 15, 91-114. doi:10.1007/BF02289195
- Hell, B., Trapmann, S., & Schuler, H. (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum [A meta-analytic investigation of subject-specific admission tests in German-speaking countries]. *Empirische Pädagogik*, 21, 251-270.
- *Hewitt, B. N., & Goldman, R. D. (1975). Occam's razor slices through the myth that college women overachieve. *Journal of Educational Psychology*, 67, 325-330. doi:10.1037/h0077010
- *Hogrebe, M. C., Ervin, L., Dwinell, P. L., & Newman, I. (1983). The moderating effects of gender and race in predicting the academic performance of college developmental students. *Educational and Psychological Measurement*, 43, 523-530. doi:10.1177/001316448304300221
- Holden, C. (1989). Court ruling rekindles controversy over SATs. *Science*, 243, 885-887. doi:10.1126/science.2919279
- *House, J. D. (1998, May). *Gender differences in prediction of graduate course performance from admissions test scores: An empirical example of statistical methods for investigating prediction bias*. Paper presented at the annual forum of the Association for Institutional Research, Minneapolis, MN. Retrieved from <http://eric.ed.gov/>
- *House, J. D., & Keeley, E. J. (1993, November). *Differential prediction of graduate student achievement from Miller Analogies Test scores*. Paper presented at the annual meeting of the Illinois Association for Institutional Research, Oakbrook Terrace, IL. Retrieved from ERIC database. (ED364605)
- Huff, K. L., Koenig, J. A., Treptau, M. M., & Sireci, S. G. (1999). Validity of MCAT scores for predicting clerkship performance of medical students grouped by sex and ethnicity. *Academic Medicine*, 74, S41-S44. doi:10.1097/00001888-199910000-00035
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
- *Jones, R. F., & Vanyur, S. (1985, April). *An investigation of gender-related test bias for the Medical College Admission Test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. Retrieved from ERIC database. (ED259024)
- *Kirchner, G. L. (1993). Gender as a moderator variable in predicting success in a Master of Arts in Teaching program. *Educational and Psychological Measurement*, 53, 155-157. doi:10.1177/0013164493053001017
- Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (Research Report No. 2008-5). New York, NY: College Board.
- Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning & Education*, 6, 51-68. doi:10.5465/AMLE.2007.24401702
- Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, 315, 1080-1081. doi:10.1126/science.1136618
- Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, 70, 340-352. doi:10.1177/0013164409344508
- *Kyei-Blankson, L. (2005). *Predictive validity, differential validity, and differential prediction of the subtests of the Medical College Admission Test* (Doctoral dissertation). Retrieved from <http://etd.ohiolink.edu/>
- Lawshe, C. H. (1983). A simplified approach to the evaluation of fairness in employee selection procedures. *Personnel Psychology*, 36, 601-608. doi:10.1111/j.1744-6570.1983.tb02237.x
- Leonard, D. K., & Jiang, J. (1999). Gender bias and the college predictions of the SATs: A cry of despair. *Research in Higher Education*, 40, 375-407. doi:10.1023/A:1018759308259
- Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136, 1123-1135. doi:10.1037/a0021276
- Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139-161. doi:10.2307/1169933
- Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 63, 507-512. doi:10.1037/0021-9010.63.4.507
- Linn, R. L. (1982). Admissions testing on trial. *American Psychologist*, 37, 279-291. doi:10.1037/0003-066X.37.3.279
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- *Luthy, T. L. (1996). *Validity and prediction bias of grade performance from Graduate Record Examination scores for students at Northern Illinois University: Age and gender considerations* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9716551)

- Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, 32, 481–498. doi:10.1016/j.intell.2004.06.008
- Lynn, R., & Kanazawa, S. (2011). A longitudinal study of sex differences in intelligence at ages 7, 11 and 16 years. *Personality and Individual Differences*, 51, 321–324. doi:10.1016/j.paid.2011.02.028
- *Lynn, R., & Mau, W. (2001). Ethnic and sex differences in the predictive validity of the Scholastic Achievement Test for college grades. *Psychological Reports*, 88, 1099–1104.
- Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (Report No. 2008–4). New York, NY: College Board.
- Meade, A. W., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods*, 12, 738–761. doi:10.1177/1094428109331487
- Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology*, 3, 192–205. doi:10.1111/j.1754-9434.2010.01223.x
- *Nauels, H., & Meyer, M. (1997). Untersuchungen zur Vorhersagekraft des TMS: Differentielle Aspekte der Studienerfolgsprognose und Testfairneß [Investigation of the predictive power of the TMS: Differential aspects of the prediction of college success and test fairness]. In G. Trost (Ed.), *Test für Medizinische Studiengänge (TMS): Studien zur Evaluation* (21. Arbeitsbericht, pp. 76–134). Bonn, Germany: ITB.
- Norborg, J. M. (1984). A warning regarding the simplified approach to the evaluation of test fairness in employee selection procedures. *Personnel Psychology*, 37, 483–486. doi:10.1111/j.1744-6570.1984.tb00524.x
- *Pape, T. E. (1992). *Selected predictors of examination for professional practice in psychology scores among graduates of Western Conservative Baptist Seminary's doctoral program in clinical psychology* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9302769)
- *Patterson, B. F., Mattern, K. D., & Kobrin, J. L. (2009). *Validity of the SAT for predicting FYGPA: 2007 SAT validity sample* (Statistical Report No. 2009–1). New York, NY: College Board.
- *Patton, T. K. (1998). *Differential prediction of college performance between gender*. Retrieved from ERIC database. (ED029407)
- *Pennock-Román, M. (1994). *College major and gender differences in the prediction of college grades* (College Board Report No. 94–2). New York, NY: College Board.
- *Qualls, A. L., & Ansley, T. N. (1995). The predictive relationship of ITBS and ITED to measures of academic success. *Educational and Psychological Measurement*, 55, 485–498. doi:10.1177/0013164495055003016
- *Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (Report No. 93–1). New York, NY: College Board.
- *Reuben, T. C. (2003). *Investigating test fairness of GRE scores for veterinary student selection* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3083156)
- Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance*, 16, 295–309. doi:10.1207/S15327043HUP1603_6
- Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, 63, 215–227. doi:10.1037/0003-066X.63.4.215
- Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2005). On interpreting research on stereotype threat and test performance. *American Psychologist*, 60, 271–272. doi:10.1037/0003-066X.60.3.271
- Sanber, S. R., & Millman, J. (1987, April). *Gender and race effects on standardized tests predictive validity: A meta-analytical study*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. Retrieved from ERIC database. (ED286914)
- Schmidt, F. L., & Hunter, J. E. (1982). Two pitfalls in assessing fairness of selection tests using the regression model. *Personnel Psychology*, 35, 601–607. doi:10.1111/j.1744-6570.1982.tb02212.x
- Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (2nd ed., pp. 257–277). New York, NY: Russell Sage Foundation.
- *Siegert, K. O. (2007). *Predicting success in graduate management doctoral programs* (GMAC Research Reports No. RR-07–10). McLean, VA: Graduate Management Admission Council.
- *Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement*, 66, 305–317. doi:10.1177/0013164405282455
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.
- Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35, 4–28. doi:10.1006/jesp.1998.1373
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613–629. doi:10.1037/0003-066X.52.6.613
- Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). Chichester, England: Wiley.
- *Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology*, 85, 710–718. doi:10.1037/0022-0663.85.4.710
- *Swinton, S. S. (1987). *The predictive validity of the restructured GRE with particular attention to older students* (GRE Board Report No. 83-25P). Princeton, NJ: Educational Testing Service.
- *Talento-Miller, E. (2008). Generalizability of GMAT validity to programs outside the U.S. *International Journal of Testing*, 8, 127–142. doi:10.1080/15305050802001193
- *Talento-Miller, E. (2009). *Validity study of non-MBA programs* (GMAC Research Reports No. RR-09–11). McLean, VA: Graduate Management Admission Council.
- *Thomas, C. L. (1973, February). *The overprediction phenomenon among black collegians: Some preliminary considerations*. Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA. Retrieved from ERIC database. (ED076679)
- *Thomas, C. L. (1979). Relative effectiveness of high school grades for predicting college grades: Sex and ability level effects. *Journal of Negro Education*, 48, 6–13. doi:10.2307/2294611
- Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, 39, 1–37. doi:10.1111/j.1745-3984.2002.tb01133.x
- *Wilson, K. M. (1982). *A study of the validity of the restructured GRE aptitude tests for predicting first-year performance in graduate study* (GRE Board Research Report No. 78-6R). Princeton, NJ: Educational Testing Service.
- Wood, J. A. (2008). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods*, 11, 79–95. doi:10.1177/1094428106296638
- *Wynne, W. D. (2003). *An investigation of ethnic and gender intercept bias in the SAT's prediction of college freshman academic performance* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3116464)

- *Young, J. W. (1994). Differential prediction of college grades by gender and by ethnicity: A replication study. *Educational and Psychological Measurement, 54*, 1022–1029. doi:10.1177/0013164494054004019
- Young, J. W., & Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (Research Report No. 2001–6). New York, NY: College Board.
- *Zeidner, M. (1987). A cross-cultural test of sex bias in the predictive validity of scholastic aptitude examinations: Some Israeli findings. *Evaluation and Program Planning, 10*, 289–295. doi:10.1016/0149-7189(87)90041-3
- Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York, NY: Routledge Falmer.

Received June 29, 2011

Revision received December 10, 2012

Accepted December 26, 2012 ■

The Internal/External Frame of Reference of Academic Self-Concept: Extension to a Foreign Language and the Role of Language of Instruction

Man K. Xu

University of Oxford and University of Cambridge

Herbert W. Marsh

University of Western Sydney, University of Oxford, and King Saud University

Kit-Tai Hau

The Chinese University of Hong Kong

Irene T. Ho

The University of Hong Kong

Alexandre J. S. Morin

University of Western Sydney

Adel S. Abduljabbar

King Saud University

The internal/external frame of reference (I/E) model (Marsh, 1986) posits that the effects of contrasting math and verbal domains of achievement are positive for matching academic self-concepts (ASCs) but negative for nonmatching ASCs (i.e., math achievement on verbal ASC; verbal achievement on math ASC). We extend the classic I/E model by contrasting the math domain with 2 verbal domains (Chinese, native language; English, foreign language) in combination with language of instruction (English or Chinese) for a sample of 1,950 Hong Kong Year 7 students. Consistent with predictions based on the Marsh and Shavelson (1985) ASC model and our extension of the I/E model, we found that native and foreign languages were not contrasted with each other in the formation of ASCs. However, achievement in both verbal domains negatively predicted math ASC, while math achievement was also negatively predicted by ASCs in both verbal domains. Support for the predictions was similar for students taught in English and Chinese languages of instruction.

Keywords: internal/external frame of reference, academic self-concept, academic achievement, language of instruction

Supplemental materials: <http://dx.doi.org/10.1037/a0031333.supp>

The purpose of the present investigation was to test extensions of the internal/external frame of reference (I/E) model of academic self-concept (ASC; Marsh, 1986; Marsh & Hau, 2004; Marsh, Martin, & Hau, 2006) in relation to the Marsh/Shavelson model of ASC (Marsh & Shavelson, 1985). Specifically, we aimed to investigate the juxtaposition of achievement and ASCs in math and

native and nonnative languages for students learning in a native language (Chinese) and a nonnative language (English) instruction environment. First, we psychometrically assessed the measurement invariance of the ASCs in math, English, and Chinese; we then evaluated the I/E model in relation to the use of native and nonnative languages of instruction (LOIs).

The Marsh/Shavelson Model and the Internal/External Frame of Reference Model of Academic Self-Concept

Shavelson, Hubner, and Stanton (1976) initially described a hierarchical, multidimensional self-concept construct in which ASC and non-ASC are two major components (see Figure 1A). Under the general ASC and non-ASCs, there are lower level self-concept factors, a series of more domain-specific self-concept factors. Based on empirical research, Marsh and Shavelson (1985) found the hierarchical nature of the self-concept construct to be weaker than had been initially thought. They proposed instead a revised model of self-concept that consisted of a higher order non-ASC factor and two higher order ASC factors—one for verbal ability and one for math (see Figure 1B for the ASC component). This was in line with the near zero correlation between verbal and math ASCs reported by Marsh and Shavelson. According to this revised model, ASCs in various school subjects form a continuum

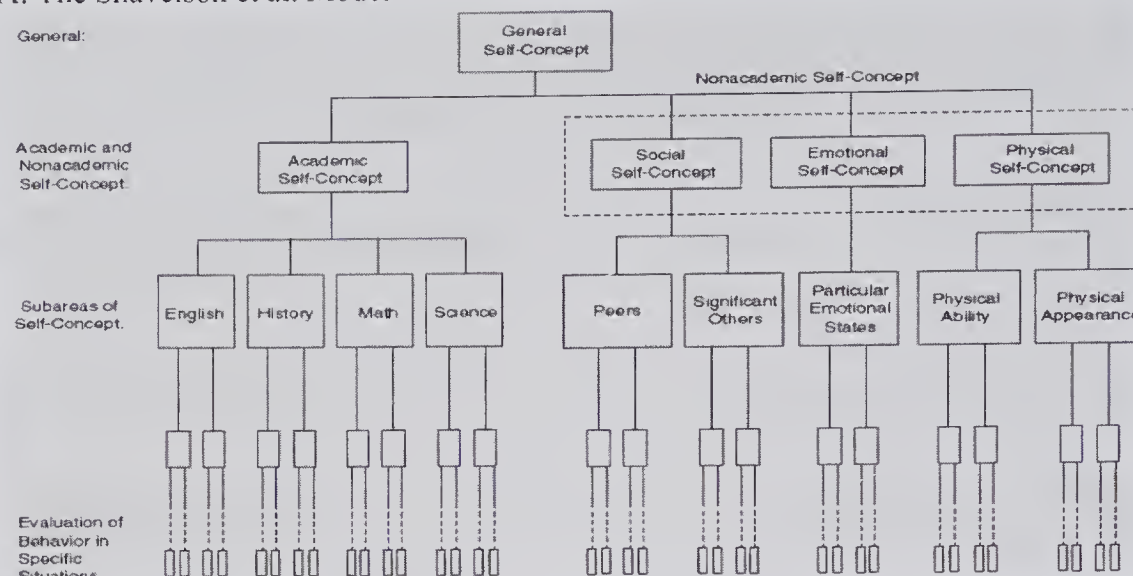
This article was published Online First February 4, 2013.

Man K. Xu, Department of Education, University of Oxford, Oxford, England, and Department of Psychiatry, University of Cambridge, Cambridge, England. Herbert W. Marsh, Centre for Positive Psychology & Education, University of Western Sydney, Sydney, New South Wales, Australia; Department of Education, University of Oxford; and Department of Psychology, King Saud University, Riyadh, Saudi Arabia. Kit-Tai Hau, Faculty of Education, The Chinese University of Hong Kong, Hong Kong. Irene T. Ho, Department of Psychology, The University of Hong Kong. Alexandre J. S. Morin, Centre for Positive Psychology & Education, University of Western Sydney. Adel S. Abduljabbar, Department of Psychology, King Saud University.

This article is based on data used in Man K. Xu's doctoral dissertation.

Correspondence concerning this article should be addressed to Man K. Xu, who is now at the Department of Psychiatry, University of Cambridge, P.O. Box 189, Addenbrooke's Hospital, Cambridge CB2 0QQ, United Kingdom. E-mail: mx212@medschl.cam.ac.uk

A: The Shavelson et al. Model



B: The Marsh/Shavelson Model

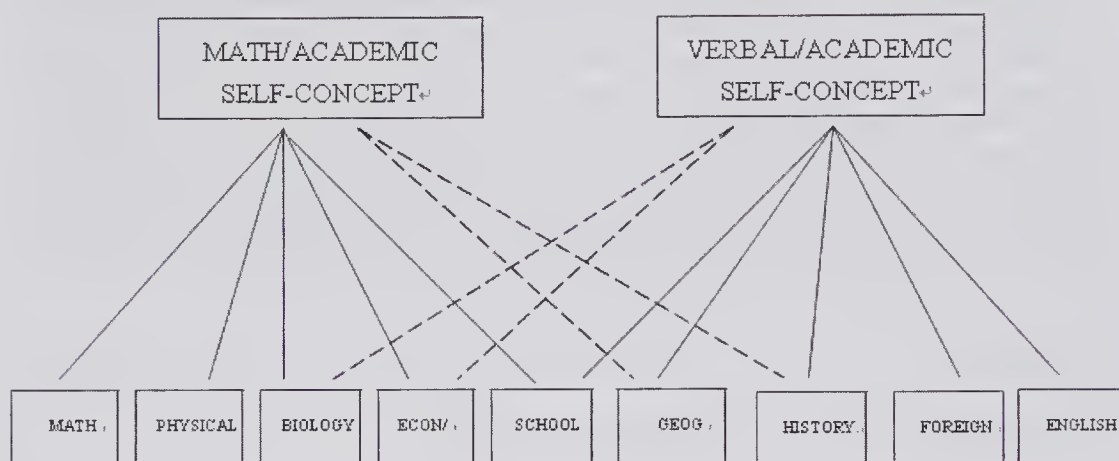


Figure 1. A: Shavelson et al.'s (1976) multidimensional, hierarchical model of SC. B: The Marsh/Shavelson (Marsh & Shavelson, 1985) model of ASC. In Figure 1A, the box enclosing non-ASC factors separates these factors from the ASC factor. The vertical lines at the bottom of the diagram indicate possible further domain-specific self-concepts. From "Self-Concept: Validation of Construct Interpretations," by R. J. Shavelson, J. J. Hubner, and G. C. Stanton, 1976, *Review of Educational Research*, 46, p. 413. Copyright 1976 by Sage. B: The boxes on the top of the hierarchy represent general math and verbal ASCs factors. The boxes in the bottom of the hierarchy represent examples of lower order components of each of the higher order ASC factors. These specific facets are presented in an order from math-oriented to verbal-oriented subjects. The dotted lines indicate that the subjects in the middle of the math/verbal continuum could load on both the math and the verbal self-concept factors, whereas the subjects at both ends of the continuum are posited to load on only one of the corresponding factors. More ASCs and hierarchies could be incorporated into the model. Econ = economics; Geog = geography; ASC = academic self-concept; SC = self-concept. From "A Multifaceted Academic Self-Concept: Its Hierarchical Structure and Its Relation to Academic Achievement," by H. W. Marsh, B. M., Byrne, and R. J. Shavelson, 1988, *Journal of Educational Psychology*, 80, p. 378. Copyright 1976 by Sage.

of ASCs, with math and verbal at opposite ends. Subjects closer to the math domain, such as physics and math, belong to the global math domain, whereas the more verbal domains such as native and foreign languages belong to the global verbal domain.

The Internal/External Frame of Reference Model of Academic Self-Concept

Although academic achievements in different academic subjects (e.g., verbal, math) are generally very positively correlated, the corresponding ASCs are nearly uncorrelated (Marsh & Shavelson,

1985). To explain this seemingly paradoxical relationship, Marsh and colleagues (Marsh, 1986; Marsh, Martin, & Hau, 2006) developed the I/E model to explain the relationship between subject-specific ASCs and achievements (Figure 2A). According to the I/E model, two underlying comparison processes or frame of reference effects are relevant to the formation of ASC: the external frame of reference process and the internal frame of reference process.

The frame of reference is the standard that students use to evaluate their ASC. According to the I/E model theory, in the external comparison process, students compare their subject specific achievement in terms of their school grades or class ranking with that of other

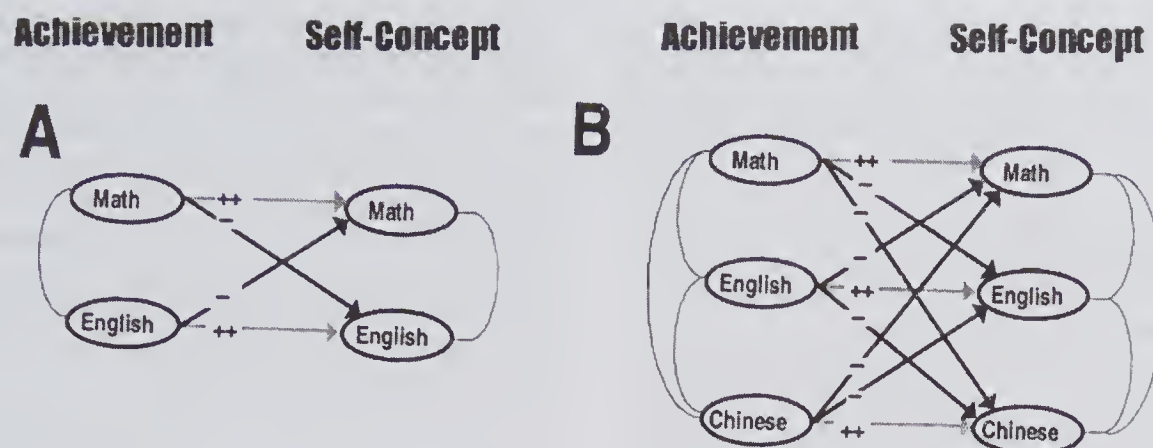


Figure 2. The I/E model and the importance of native and nonnative languages. A: I/E model predictions based on the traditional two subjects. B: An extended I/E model taking into account native and nonnative languages and math. The horizontal lines represent substantial and positive effects (++) from achievement to self-concept. The cross paths represent smaller and negative effects (–) from achievement to self-concept. I/E = internal/external frame of reference. From “Extension of the Internal/External Frame of Reference Model of Self-Concept Formation: Importance of Native and Nonnative Languages for Chinese Students,” by H. W. Marsh, C. K. Kong, and K. T. Hau, 2001, *Journal of Educational Psychology*, 93, p. 544. Copyright 1976 by Sage.

students in the same school or classroom and use this information to form their own ASC in that subject. In this respect, the formation of ASC is likely to be a function of the students' relative ranking in their school or classroom. Higher achievement relative to others leads to higher ASC. This process is presented as the horizontal paths in the I/E model (see Figure 2A).

For the internal comparison process, students use information from achievements in all their school subjects in forming their ASC in each subject. For example, if verbal achievement is higher than math achievement, verbal ASC is likely to be higher than math ASC. In the I/E model, these predictions are represented by the negative cross paths (see Figure 2A). Hence, after controlling for the paths from achievement to ASC in the matching domains, verbal achievement is negatively predicted math ASC, and vice versa. Within this pattern of internal comparisons, a student who has higher math achievement compared to his or her verbal achievement may have a reasonable math ASC even if the student is weak in both math and verbal domains in comparison to other students. The combination of the external comparison process (predicting positive associations between achievements and matching ASCs) and internal comparison process (predicting negative associations between achievements and nonmatching ASCs) is consistent with the near-zero correlation between math and verbal self-concepts.

It is important to emphasize that the I/E model is based on the paths from multiple achievement domains to multiple ASC domains. For example, math achievement is not predicted to be negatively correlated with math ASC without controlling for verbal achievement. Although the simple correlations are relevant and of interest in relation to demonstrating the extreme domain specificity of ASCs—particularly in relation to corresponding areas of achievement—they do not provide direct tests of the I/E model.

The I/E model provides strong support for the discriminant validity of the domain-specific ASCs, showing that each ASC has a distinct pattern of relations with corresponding measures of academic achievement. For example, English ASC is predicted positively by English achievement but negatively by math achieve-

ment, and math ASC is positively predicted by math achievement but negatively predicted by English achievement. This is in agreement with previous support for the multidimensionality of ASC that the measures of verbal and math ASC are distinct constructs. A general measure for both would mask the unique relationship specific to math and English ASCs.

Indeed, the ASCs are much more differentiated than the corresponding measures of achievement. After controlling for the effect of math (verbal) achievement on math (verbal) ASC, the effect of verbal (math) achievement on math (verbal) ASC is posited to be negative. This explains in part why people tend to think of themselves as either a “math person” or a “verbal person.” Even very capable students typically perceive differences in their ability levels in different school subjects and thus might have an average or even below-average ASC in their weakest academic subject. Likewise, even academically less able students may have an average or even above-average ASC in their best performing school subject (also see Möller, Streblov, & Pohlmann, 2006).

Basing the theoretical framework on the hierarchical and multidimensional characteristics of the ASC construct, the I/E model has been widely explored and supported in the empirical literature, through different ASC instruments (e.g., Marsh, Byrne, & Shavelson, 1988; Tay, Licht, & Tate, 1995), over time (e.g., Köller, Klemmert, Möller, & Baumert, 1999; Marsh & Köller, 2004; Marsh, Kong, & Hau, 2001; Marsh & Yeung, 1998; Möller & Köller, 2001b; Möller, Retelsdorf, Köller, & Marsh, 2011), and across samples of high ability (Mui, Yeung, Low, & Jin, 2000; Plucker & Stocking, 2001) and low ability (Möller, Streblov, & Pohlmann, 2009) students. Evidence was also found in experimental studies (Möller & Köller, 2001a; Pohlmann & Möller, 2009) and cross-cultural studies (Chiu, 2008, 2012; Marsh & Hau, 2004).

A recent meta-analysis based on 69 data sets ($n = 125,308$) also offers remarkably strong support for the I/E model (Möller, Pohlmann, Köller, & Marsh, 2009), used to evaluate the effects of math and verbal achievements on ASCs. Math and verbal achievement were much more highly correlated than the corresponding math and verbal ASCs (0.67 vs. 0.10). The I/E model analysis showed

that controlling for the effect of verbal achievement to verbal ASC (0.49) and math achievement to math ASC (0.61), verbal achievement negatively predicted math ASC (-0.27), and math achievement negatively predicted verbal ASC (-0.21). The results were robust in relation to age, gender, and country groups.

Extension of the I/E Model: Assimilation and Contrast Along a Continuum of ASCs

The I/E model has been studied almost exclusively for the qualitatively distinct math and verbal (the native language) domains. This body of literature mainly comes from the original Marsh/Shavelson (Marsh & Shavelson, 1985) model (see Figure 1B) that hypothesizes a continuum of ASC domains (Marsh, Byrne, & Shavelson, 1988). The Marsh/Shavelson model suggests that ASCs can be ordered along a math-verbal continuum, with math and native language as the two end-points of this continuum (Marsh, 1990, 1992). Math and verbal ASCs, which are at the opposite ends of the continuum, are the most contrasted ASCs, and there is much support for this hypothesis in the form of the I/E model. We use the term contrast in the sense that, consistent with predictions from the classic I/E model (Marsh, 1986), the effect from achievement in a domain to ASC in a nonmatching domain is negative.

Less studied is whether the I/E model also holds for ASCs that are closer together on the continuum (e.g., math vs. physics; native vs. native languages). Following logically from the theoretical rationale for the I/E model, coupled with the Marsh/Shavelson (Marsh & Shavelson, 1985) model and the continuum of ASCs (e.g., Figures 1A & 1B), we predict that the contrast effect will be stronger, the further apart the two domains are on the ASC continuum. That is, the negative effect of achievement in one domain will be more negative for domains that are further apart. This leaves open the intriguing possibility that there might be assimilation for domains that are closer together on the continuum. That is, for closely related domains, the effect of achievement in one domain on ASC in a different domain might be positive (assimilation) rather than negative (contrast). Making a similar point as a direction for further research, Möller, Pohlmann, Koller & Marsh (2009) asked,

Would students see physics and mathematics as sufficiently distinct that better performances in one would lead to poorer self-concepts in the other (a contrast effect like that posited in the I/E model based on the math and verbal domains), or would the two be seen as sufficiently similar so that better performance in one would lead to better self-concepts in the other (an assimilation effect)? Although clearly beyond the scope of the present investigation, this is a potentially important extension of the I/E model. (p. 1159)

Although this is apparently a new extension of the classic I/E model (but see Marsh & Yeung, 2001; Möller, Pohlmann, et al., 2009; Möller, Streblow, Pohlmann, & Köller, 2006), there is some empirical support for it.

The strongest evidence for our extension of the I/E model comes from the German study by Möller et al. (2006), who juxtaposed achievement and ASCs in math, physics, German (native language), and English (foreign language) in a sample of seventh to 10th grade students. Consistently with classical I/E predictions, they found the expected negative paths from contrasting domains (German and English achievements to math and physics ASCs;

math and physics achievements to German and English ASCs). However, there was assimilation for math and physics—small positive effects from physics achievement to math ASC and from math achievement to physics ASC. Although the paths from English to German and German to English were all consistently small, they clearly did not support the strong contrast effects predicted by the classic I/E model. Möller et al. (2006) suggested that students apparently perceive native and foreign languages as being more distinct than math and science and suggested that this might be due to different teaching strategies in native and nonnative languages that accentuate their distinctiveness.

Particularly relevant to the present investigation, Marsh and Yeung (2001; a reanalysis of Bong, 1998) extended the I/E model juxtaposing to math, Spanish, and English achievements and ASCs for 383 American students (Grades 11 and 12), a majority of whom were of Spanish descent (16% White, 6% African American, 55% Hispanic, 20% Asian, 2% Native American). Based on five achievement test scores they posited three achievement factors (verbal, math, and Spanish) that were related to six ASCs (global verbal, English, Spanish, history, global math, algebra, geometry, and chemistry). Analyses based only on the global verbal and math ASCs and on corresponding achievement scores showed clear support for classic I/E predictions. When expanded to include all the achievement and ASC scores, Spanish was distinct from the other domains in relation to both achievement and ASC. Verbal achievement had positive effects on English, history, and global verbal ASCs, but negative effects on Spanish, algebra, geometry, and chemistry ASCs, while math achievement had negative effects on English, history, verbal, and Spanish ASCs but positive effects on algebra, geometry, global math, and chemistry ASCs. Spanish achievement had small negative effects on all ASC scales other than Spanish ASC. Although this is largely consistent with our proposed extension of the I/E model, the results suggest that Spanish in this study was quite distinct from domains near the verbal end of the continuum as well as from the math end of the continuum. In this study, Spanish, the native language for many participants, was found to negatively predict competence evaluation in other school subjects. However, given that Spanish was the native language for nearly half of the students and a foreign language for the other half, the results may have been idiosyncratic to this sample. Further, the sample size was not sufficient to test the generalizability of results across different ethnic groups.

Several previous I/E studies, based on Hong Kong students, are also particularly relevant to the present investigation. In a large longitudinal study (Marsh et al., 2001) of Hong Kong high school students (Grade 6 to Grade 10), ASCs and achievement in three domains (math, Chinese, English) showed contrast effects: negative paths from achievement scores to nonmatching ASC in all domains (see also Yeung, Lee, & Wong, 2001). However, even though there were negative effects of English achievement on Chinese ASC and of Chinese achievement on English ASC, the effects of Chinese achievement on English ASC were smaller (or nonsignificant) than paths relating Chinese or English achievement to math ASC. This indicates a much weaker internal frame of reference effect from Chinese achievement to English ASC. Furthermore, based on a sample of university students, Yeung and Lau (1998) found that support for the I/E model predicted contrast effects for relations between the English (nonnative language) and math domains but not for relations between the Chinese (native

language) and math domains. However, a limitation of this study was that they did not juxtapose the three school subjects in one model; rather, they tested the traditional I/E model based on two domains at a time.

In summary, for academic domains at opposite ends of the ASC continuum, I/E research consistently shows support for classic I/E predictions—contrasting, negative effects of achievement on ASCs in nonmatching, distinct domains. Nevertheless, for domains closer together (e.g., math and science, or native and non-native language), the results are mixed, and marked by much weaker contrast effects, no significant effects, or even assimilation effects between achievement and ASC in nonmatching domains.

Native and Nonnative Language and Language of Instruction (LOI)

Of particular relevance to the present investigation, there is no clear consensus as to whether native and nonnative languages are perceived as two similar verbal domains or as distinct domains. A related issue is the language of instruction (LOI). In Hong Kong, the juxtaposition of English-LOI versus Chinese-LOI is a critical issue with important substantive and policy–practice implications. Hong Kong has a long history of bilingual education (Evans, 2011), with both English and Chinese LOIs in Hong Kong secondary schools (Grades 7 to 12). The LOI policy in Hong Kong is a type of late immersion, where the LOI changes from Chinese in most primary schools to English in about a quarter of the more prestigious secondary schools (Marsh, Hau, & Kong, 2000). On the one hand, the Marsh et al. (2001) study of secondary school students found some support for the classic I/E model contrast predictions among English and Chinese subjects. On the other hand, the English language is qualitatively very different from the Chinese language, so a dimensional comparison aspect could still exist between English and Chinese, particularly when English is taught as a separate subject in Chinese LOI schools. Conversely, English and Chinese might be seen as complementary domains when Chinese students are immersed in English LOI schools.

In the present investigation, we explore the implications of the LOI in relation to students' perception of their academic competence and the I/E model. For instance, following suggestions by Möller, Pohlmann, et al. (2009), is it possible that when English is the LOI for most academic subjects, students would perceive English and Chinese as more similar (two verbal subjects) rather than more distinct foreign-language and native-language subjects? If so, English LOI students might regard the combination of English and Chinese as a single basis of comparison for evaluating their ASCs in other, more distinct subjects such as math. Do students attending English LOI schools perceive their English and Chinese competence and ASCs similarly to the students attending Chinese LOI schools? Are the processes posited in the I/E model, relating achievement and ASC (in this extension including English and Chinese as well as mathematics) similar for students in LOI-English and LOI-Chinese schools? To date, no study has examined in detail a LOI effect on either the measurement or structural aspects of the I/E model.

The Present Investigation

Extending the traditional social comparison basis of ASC formation that students compare their accomplishments with classmates as one basis of forming their self-concept, the I/E model posits that as an internal frame of reference process, students also compare their achievements in different school subjects. The model generates the seemingly paradoxical prediction that high achievement in one school subject will have a negative effect on ASC in a contrasting domain. However, the considerable body of support for this prediction has been limited primarily to the math domain and the verbal domain represented by the native language—two school subjects that are maximally differentiated along the math–verbal continuum posited in the Marsh/Shavelson (Marsh & Shavelson, 1985) model of self-concept. Hence, the purpose of the present investigation is to extend the empirical and theoretical basis of the well-established I/E model based on math and verbal constructs and include LOI that involves a nonnative language. The overarching research question is whether the nonnative LOI acts as a verbal-like school subject such that achievement in this subject is contrasted with math achievement but not achievement in the native language or whether it is contrasted with achievements in both math and native language.

We evaluate this overarching research question with a series of latent-variable statistical models that extend those used in traditional I/E studies. More specifically, we evaluate

- (a) the structure of ASC responses, comparing first-order models in which ASCs in math, English, and Chinese are each posited as separate constructs, and a second-order model in which English and Chinese ASCs are postulated to belong to a single domain representing a general verbal ASC, and the generalizability of the ASC structure across LOI groups;
- (b) the traditional I/E models involving two academic subjects, in math versus English, math versus Chinese, and English versus Chinese;
- (c) the juxtaposition of the three subjects through a first-order I/E model where the ASCs in math, Chinese, and English are predicted by achievements both in their matching and nonmatching domains;
- (d) the juxtaposition of the three subjects through higher order factor I/E models. In this model, a verbal ASC incorporating English and Chinese ASCs is predicted by both the math achievement and a corresponding verbal achievement factor based on Chinese and English achievements. Similarly math ASC is predicted by the corresponding math achievement and nonmatching verbal achievement; and
- (e) the invariance tests of the I/E model across LOIs. This step systematically assesses the extent to which the I/E models differ across LOI groups, in particular whether the path coefficients between ASCs and achievements are equivalent in both groups. Due to space limitations, the full results of Research Question 1 are presented in the sup-

plemental material (Part II) and only briefly treated as preliminary analysis in the Results section.

Method

Sample

The data consisted of a sample of Hong Kong secondary school students ($n = 1,950$; 47.3% boys, 52.7% girls) from the end of the school year in Grade 7 (47 intact classes, 12 schools) who were asked about motivational aspects of school learning. The schools were sampled from various districts and were broadly differentiated in terms of academic strength. Of the 12 schools, four were from above-average school ability bands, four were from average-ability bands, and four were from below-average school ability bands.

Language of Instruction

In Hong Kong, essentially all mainstream public primary schools have Chinese LOI, and these students have limited ability in English. All students are able to apply to any Hong Kong public high school, although English-LOI schools typically are seen as more desirable and are able to select academically more able students, based on standardized tests completed by all students at the end of primary school. These include standardized English tests—one basis of selection for students to attend English-LOI schools. A third of the students were from English-LOI schools, the rest were from Chinese-LOI schools. All of the participants were between 11 and 16 years old ($M = 12.0$, $SD = 0.89$): 86% were 12–14 years old, 60% were 12 years old, 25% were 13 years old, and 8% were 14 years old. Hence, the majority of the students were within the normal age range for this academic grade. For Chinese LOI students, 41.81%, 43.29%, and 14.90% were from the high, medium, and low ability bands, respectively. For English LOI students, 34.53% and 65.47% were from the high and middle ability bands, respectively; no English LOI students were from low ability band. The mean ages of English- and Chinese-LOI students were similar ($M_s = 12.37$ and 12.65 , respectively).

Data were collected at the end of Grade 7, the 1st year of secondary schooling, so the students who were in English LOI schools would have had 1 year of English LOI experience.

Measures

The participants completed questionnaires measuring ASCs in math, Chinese and English at the end of Grade 7. The ASC instrument consisted of three items for each of the school subjects, asking students the extent to which they agreed with the statements on a 6-point Likert response scale ranging from *strongly disagree* to *strongly agree* (e.g., I do well in tests in this subject.). The Cronbach's alpha coefficients of reliability were .80 for Chinese, .83 for English, and .83 for math.

Two sets of achievement measures were considered: a standardized achievement test taken by all the students in Hong Kong (in July, when they were still in Grade 6) prior to their entry into the 1st year of secondary schooling (Grade 7, in September) and their school marks, the overall grades from students' end-of-year exams in Grade 7. The final achievement scale used in the present

investigation was derived from school marks moderated by the standardized achievement test results so that the final achievement scale was comparable across schools and classrooms (see supplemental materials, Part I, for detailed explanation; see also Marsh et al., 2001; and Xu, 2010).

Statistical Analysis

We used structural equation models (SEM; e.g., Byrne, 2001; Schumacker & Lomax, 2004; Tabachnick & Fidell, 2006) for statistical analysis. The fit of the models was evaluated by a range of recommended fit indices (Marsh, Balla, & Hau, 1996; Marsh, Balla, & McDonald, 1988), which included the Tucker-Lewis index (TLI), the root-mean-square error of approximation (RMSEA), the comparative fit index (CFI), the χ^2 statistic, and the standardized root-mean-square residual (SRMR). CFIs and TLIs greater than .95 indicate an acceptable model fit, whereas RMSEAs less than .06 indicate good fit. Multiple group analysis was used to assess the measurement invariance of the ASCs and their structural relations with academic achievement across groups of LOI (see supplemental material for description of this method).

Data were gathered using survey items of parallel wording such as "I do well in tests in math" and "I do well in tests in English." Without taking into account the method effect associated with parallel wording, the model fit is likely to be less than adequate and the parameter estimates might be biased, leading to potentially invalid interpretations of the results (Marsh & Hau, 1996). Following suggestions by Marsh and Hau (1996), correlated uniqueness was specified in the models as a priori.

Since the data are hierarchical (students nested within classes), a complex design correction was applied to the model estimation in the Mplus software in conjunction with SEM models through TYPE = COMPLEX in Mplus software (L. K. Muthén & Muthén, 2007). This complex design function takes cluster-sampling into account in estimates of standard errors (e.g., B. O. Muthén & Satorra, 1995; Stapleton, 2006).

The amount of missing data was small; 1.28% for ASC responses, and 0.31% for the achievement data. The Multiple Imputation (Collins, Schafer, & Kam, 2001; Schafer & Graham, 2002) method was used to counter the missing data problem. Ten complete sets of data were generated with the software package NORM (Schafer, 1999), then analyzed using the robust maximum likelihood estimator in Mplus software.

Results

Preliminary Analyses: The Structure of the ASC and Its Generalizability Over LOIs (Supplemental Material, Part II)

To empirically test extensions of the Marsh/Shavelson (Marsh & Shavelson, 1985) model in native and nonnative languages, we first tested a confirmatory factor analysis (CFA) model with achievement and ASCs in math, English and Chinese, modeled as first-order factors (Model TGS1, Table S1, Figure S1A, supplemental material). Correlations among the achievement and ASC factors (Model TGS1, Table S2, supplemental material) showed that math achievement was positively correlated with math ASC (0.39) and English ASC (0.11) but not significantly correlated with

Chinese ASC; English achievement was positively correlated with English ASC (0.42) but not significantly correlated with Chinese or math ASC; Chinese achievement was positively correlated with both English ASC (0.22) and Chinese ASC (0.19) but was not significantly correlated with math ASC. While achievement measures in all three subjects were highly correlated (0.63 to 0.72), math ASC was only weakly correlated with English ASC (0.19) and Chinese ASC (0.24). On the other hand, English and Chinese ASCs were substantially correlated with each other (0.45).

Next, we tested higher order CFA models with a higher order verbal ASC incorporating both English and Chinese ASCs, as well as a verbal achievement factor consisting of English and Chinese achievement (Model TGS2, Table S1, Figure S1 B, supplemental material). In this model, math achievement was substantially correlated with math ASC (0.40) but only moderately correlated with verbal ASC (0.11; Table S2, Model TGS2). Similarly, verbal achievement was correlated with verbal ASC (0.44) but was not significantly correlated with math ASC. While verbal and math achievements were highly correlated (0.75), math and verbal ASCs were only weakly correlated (0.26).

We compared both versions of the models' specifications and demonstrated that the English and Chinese ASCs could form a single, higher order verbal ASC and also that English and Chinese achievements could be modeled through with a single verbal achievement factor (Figure S1, Table S1, supplemental material). We then proceeded to multiple group CFA analysis based on Model TGS1 and confirmed that the measurement properties of ASCs as well as their correlations with achievements were invariant for students in different LOIs (Table S1, supplemental material). These results provide a basis for us to explore further aims in the present investigation.

The I/E Models in Math and English, in Math and Chinese, and in Chinese and English

We begin with initial tests of the I/E model of relations between achievement and ASC based on different pairs of school subjects. The results for the I/E model for math and English (see Figure 3A, Model TG1, Table 1), for math and Chinese (Model TG2, Figure 3B, Table 1), and for Chinese and English (Model TG3 Figure 3C, Table 1) showed that in general there was a high positive relation between achievement and ASC in the matching domains but moderate negative relations between achievement and ASC in nonmatching domains. However, the relationship between nonmatching domains in the English–Chinese (-0.17 , -0.15) I/E model was somewhat weaker compared to the relationship found in the English–math (-0.30 , -0.27) and the Chinese–math (-0.34 , -0.22) models. The weak cross paths and high correlations between Chinese and English ASCs suggest a weak internal comparison mechanism with regard to self-perceived abilities in English and Chinese. This is in line with the finding that a second-order verbal factor could incorporate both the English and the Chinese ASC first-order factors (supplemental material, Part II).

The I/E Model in Math, English, and Chinese—A First-Order I/E Model

Model TG4 (see Figure 4A, Table 1) included all three academic subjects in a single model. Model TG4 was well-defined, in that all factor loadings were substantial and the fit indices were excellent

(e.g., CFI = 0.993). There were high positive correlations between the math and English achievements (0.65), math and Chinese achievement (0.63), and English and Chinese achievement (0.72) but substantially smaller correlations between ASCs in these subjects (0.32 for math and English, 0.36 for Chinese and math and 0.50 for English and Chinese). The paths from math achievement to math ASC (0.66), from English achievement to English ASC (0.63), and from Chinese achievement to Chinese ASC (0.36) were all significant and positive. Although differing in size, all of the cross paths were negative. The paths from math achievement to English ASC (-0.25) and from math achievement to Chinese ASC (-0.19) were both significant and negative, controlling for the effects of matching domain achievements on ASCs. English achievement negatively predicted math ASC (-0.16), but its negative effect on Chinese ASC (-0.08) was not statistically significant. Similarly, for Chinese achievement, the path leading from Chinese achievement to math ASC (-0.26) was negative and significant, but the negative effect of Chinese achievement on English ASC was not statistically significant (-0.08). This shows that the I/E pattern of relations involving English and Chinese was no longer statistically significant once all three subjects were posited simultaneously in one model.

In summary, math achievement negatively predicted both English and Chinese ASCs, and Chinese achievement and English achievement both negatively predicted math ASC. However, the paths leading from English achievement to Chinese ASC and from Chinese achievement to English ASC were not statistically significant. In the next step, we combine achievement and ASCs in English and Chinese into higher order factors, in order to implement an alternative test of the I/E model.

The I/E Model in Math, Chinese, and English—A Higher Order Model

In Model TG5 (see Figure 4B, Table 1), we posited a model including a second-order verbal ASC factor with English and Chinese ASCs as first-order factors. In this model, a verbal achievement factor was constructed using English and Chinese achievement as indicators. Math achievement positively predicted math ASC (0.77), and verbal achievement positively predicted verbal ASC (0.82). The correlation between math and verbal achievement (0.75) was higher than the correlation between math and verbal ASCs (0.56). Most importantly, the cross paths from nonmatching domains were substantial and negative (-0.50) from math achievement to verbal ASC, and -0.50 from verbal achievement to math ASC. This supports the I/E model theory and a multidimensional verbal ASC construct in that, while the lower order constructs were subject-specific, they formed a higher order general verbal factor that negatively predicted math ASC.

Both Model TG4 and Model TG5 showed excellent fit indices and a substantively consistent pattern of I/E model relations. Model TG4 did fit the data slightly better in comparison to Model TG5, but Model TG5 provided a more parsimonious representation of the results and offered a clearer interpretation of the I/E model relations with regard to the math and verbal domains. Importantly, both models are consistent with the conclusion that students did not use English as an internal frame of reference in the evaluation of their Chinese skills, and vice-versa. In this sense, the substantive interpretations based on the two models are similar. The next step

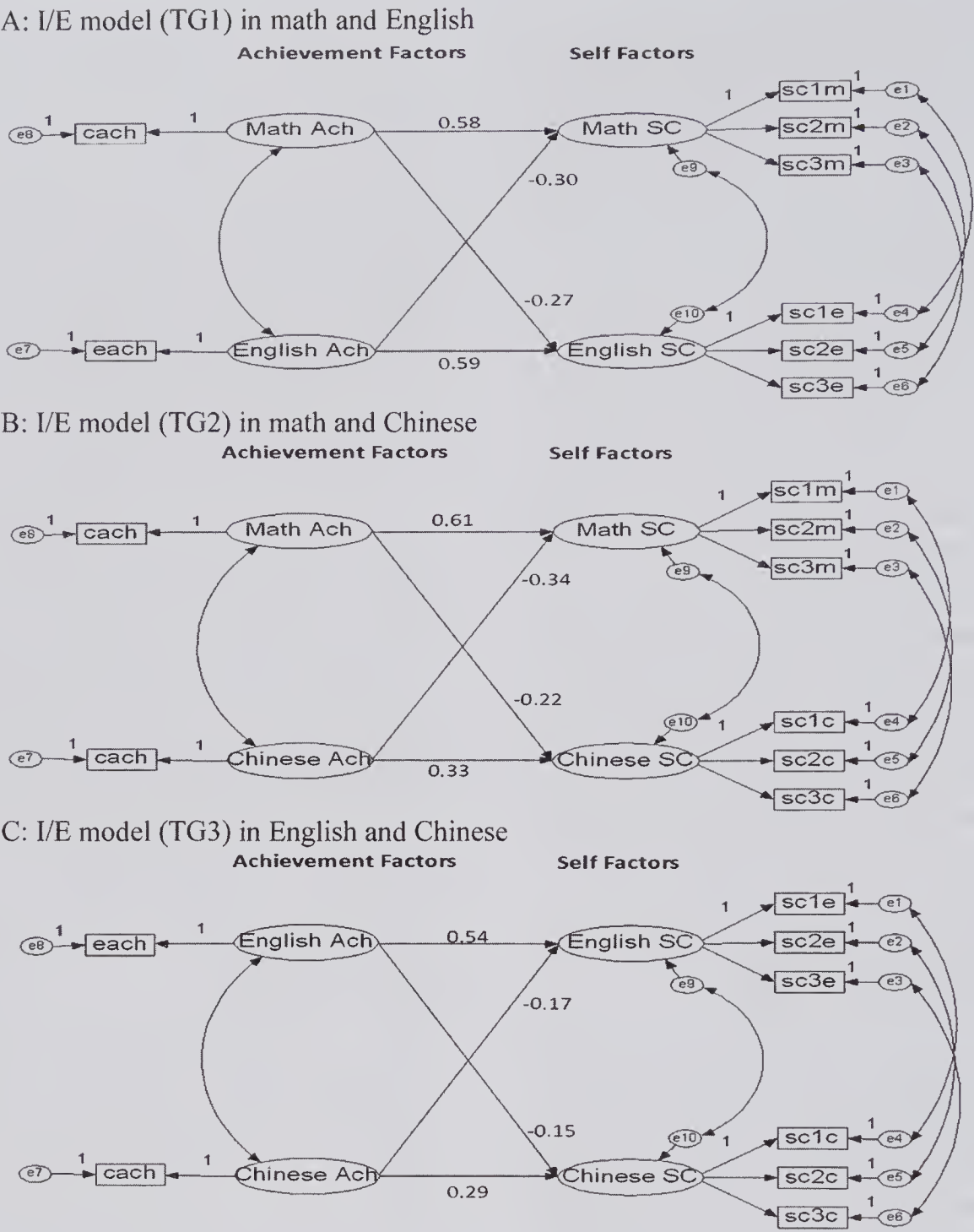


Figure 3. Traditional two-subject I/E models. Separate I/E models of math and English (A, Model TG1), math, and Chinese (B, Model TG2) and Chinese and English (C, Model TG3). Single-indicator factors were defined such that the standardized factor loading was 1 and uniqueness was 0. The covariances for self-concept item residuals were correlated uniqueness for the parallel worded items. The path coefficients were statistically significant at $p < .05$. I/E = internal/external frame of reference; Ach = achievement; TG = model based on a single group; SC = self-concept; cach = Chinese achievement; mach = math achievement; each = English achievement; m = math; c = Chinese; e = English.

tested whether Model TG4 differed across LOI. We chose Model TG4 as it would allow us to examine specific paths between achievement and ASC in all three subjects.

I/E Model in Math, English, and Chinese, and the Effect of the LOI

We now move to the research question as to whether the I/E model is invariant for LOI groups: Chinese students in English

immersion schools compared to those in native Chinese language schools. We accomplish this by considering multiple group tests of the invariance of the factor structure over the two groups. Although we test the invariance of a variety of different parameters, the most critical ones are the paths leading from achievement to ASCs that are central to the I/E model. Models MG4_1 to MG4_5 (see Table 1) were constructed to test the group invariance of Model TG4 in five models across two LOI groups. The strategy

Table 1
Summary of Goodness of Fit for SEM Models TG1–MG4_10

Model	χ^2	df	CFI	TLI	RMSEA	SRMR	Description
SEM I/E models of 2 subjects							
TG1	60.533	13	0.992	0.982	0.043	0.017	Figure 3A: I/E math, English
TG2	41.533	13	0.995	0.990	0.034	0.015	Figure 3B: I/E math, Chinese
TG3	46.813	13	0.995	0.988	0.037	0.017	Figure 3C: I/E English, Chinese
SEM FO and HO I/E models in math, English, and Chinese							
TG4	105.865	33	0.993	0.987	0.034	0.019	Figure 4A: I/E math, English and Chinese
TG5	370.994	38	0.969	0.947	0.067	0.040	Figure 4B: I/E math, HO vsc, HO vach
Multiple-group FO SEM I/E models in math, English, and Chinese (TG4, Figure 4A)							
MG4_1	159.327	66	0.990	0.980	0.038	0.022	SEM INV = none; Free = FL, PC, FV, FC, Uniq., CU, Inter (FMns = 0)
MG4_2	174.820	72	0.989	0.980	0.038	0.024	SEM INV = FL; Free = PC, FV, FC, Uniq., CU, Inter (FMns = 0)
MG4_3	207.905	81	0.986	0.978	0.040	0.034	SEM INV = FL, PC; Free = FV, FC, Uniq., CU, Inter (FMns = 0)
MG4_4	203.577	78	0.987	0.977	0.041	0.031	SEM INV = FL, PC–; Free = PC+, FV, FC, Uniq., CU, Inter (FMns = 0)
MG4_5	185.529	75	0.988	0.979	0.039	0.030	SEM INV = FL, PC+; Free = PC–, FV, FC, Uniq., CU, Inter (FMns = 0)

Note. Models labeled TG are based on a single group, whereas MG refers to models with multiple groups. SEM = structural equation modeling; I/E = internal/external frame of reference; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; FO = first order; HO = higher order; sc = self-concept; ach = achievement; vsc = verbal self-concept; vach = verbal achievement. For multiple-group invariance models, INV = parameters constrained to be invariant across the multiple groups; FL = factor loadings; PC = path coefficients; FV = factor variances; Uniq = item uniqueness; FC = factor covariances; CU = correlated uniquenesses; Inter = item intercepts; FMn = Factor means; PC– = path coefficients hypothesized to be negative; PC+ = path coefficients hypothesized to be positive.

here is similar to the multiple-group CFA invariance tests in the supplemental material. We first examine the factorial invariance of the LOI groups in Models MG4_1 and MG4_2 (see Table 1), then we move to the structural invariance, focusing on the path coefficients posited in the I/E model (MG4_3 to MG4_5).

Factorial invariance. Through Model MG4_1 and Model MG4_2, we firstly established factorial invariance across LOI groups—the invariance of factor loadings relating indicators and latent constructs. Support of factor loading invariance was demonstrated by the minimal change in the fit indices when factor loadings were constrained to be equal across two groups (see discussion of fit indices in the Method section and supplemental material): $\Delta\text{CFI} = -0.001$, $\Delta\text{TLI} = 0.000$, $\Delta\text{RMSEA} = 0.000$; thus, loading invariance was established across the LOI groups.

Structural invariance. In Models MG4_3 to MG4_5, invariance of the path coefficients between academic achievement and ASCs was tested. In Model MG4_3, all path coefficients were constrained to be the same across the two groups. Compared with Model MG4_2, where all path coefficients were free, constraining nine path coefficients only led to negligible decreases in fit indices (see earlier discussion of fit indices): $\Delta\text{CFI} = -0.003$, $\Delta\text{TLI} = -0.002$, and $\Delta\text{RMSEA} = 0.002$.

In Models MG4_4 and MG4_5, regression paths were grouped in terms of the a priori predictions of the direction of effects and tested separately in relation to the I/E model theory. In Model MG4_4 the three matching-domain path coefficients were freed

but all six cross-path coefficients were constrained to be equal across LOI groups. Model MG4_5 constrained three matching domain path coefficients. Compared to Model MG4_2 where all coefficients were free, both Models MG4_4 and MG4_5 showed only very slight decreases in CFI and TLI.

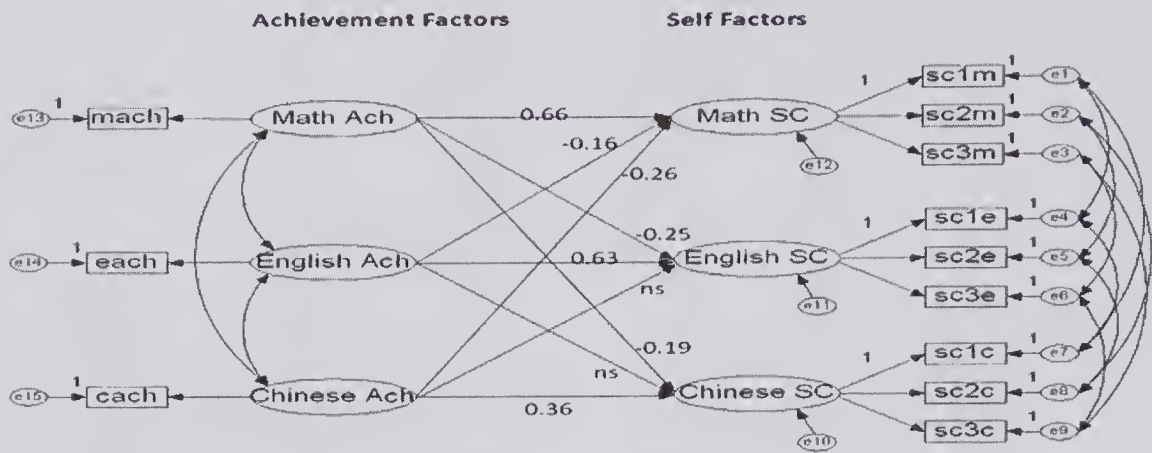
Based on the results from the multiple-group invariance tests, both the factorial and the structural models of ASC and achievements were shown to be invariant across LOI. Thus, we conclude that the I/E model generalizes well over the two LOIs, providing further strong support for the generalizability of the I/E model.

Discussion

Based on a sample of Hong Kong secondary school students under English and Chinese LOIs, using ASC and academic achievement measures in math, English, and Chinese, the present investigation extends the traditional math/verbal I/E model and integrates more fully the contrast and assimilation concepts proposed by the Marsh/Shavelson (Marsh & Shavelson, 1985) model of ASC. More specifically, the present study investigated and concluded the following research questions:

1. To clarify the structure of native and nonnative ASCs in relation to math ASC, alternative latent variable models were specified. We firstly looked at math, English, and Chinese ASCs in a first-order CFA model, then examined a

A: I/E Model (TG4) in Math, English and Chinese



B: I/E Model (TG5) in Math, Higher-order Verbal Self-concept, and Higher-order Verbal Achievement in English and Chinese

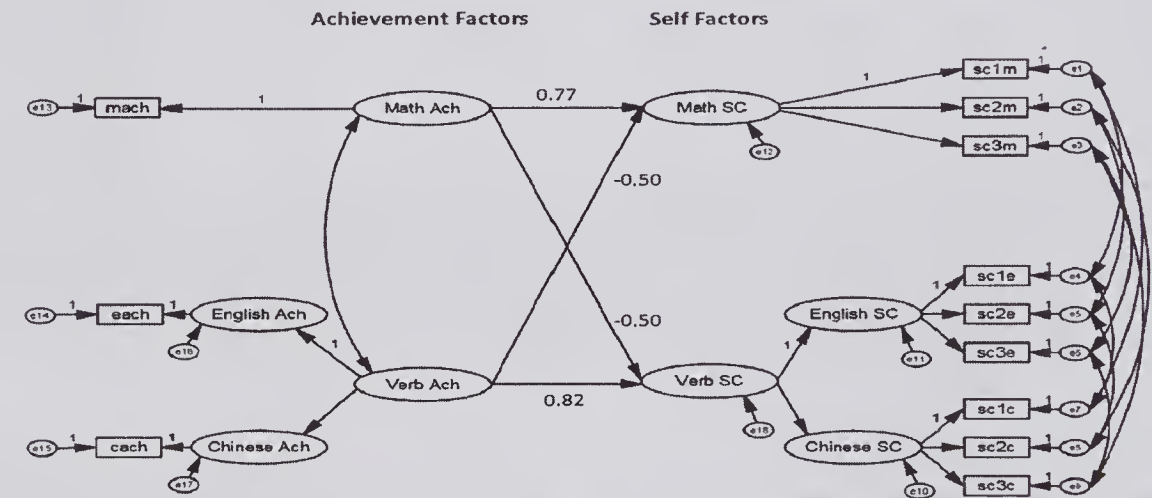


Figure 4. I/E models in three school subjects (parameter estimates included). I/E model of first-order achievement and self-concept factors in math, English, and Chinese (A, Model TG4), and I/E model of first-order achievement and self-concept factors in math and second-order achievement and self-concept factors in verbal, incorporating first-order factors in Chinese and English (B, Model TG5). Single-indicator factors were defined such that the standardized factor loading was 1 and uniqueness was 0. The covariances for self-concept item residuals were correlated uniqueness for the parallel worded items. The path coefficients were statistically significant at $p < .05$ except when specified to be *ns* ($p < .05$). I/E = internal/external frame of reference; TG = model based on a single group; mach = math achievement; ach = achievement; SC = self-concept; each = English achievement; cach = Chinese achievement; verb = verbal; m = math; c = Chinese; e = English.

second-order model where ASCs in English and Chinese were combined in a second-order factor to contrast with math ASC. Results supported the notion that English and Chinese ASCs represent a unified verbal domain. That is support that the ASC constructs are both multidimensional and subject-domain-specific, as predicted by the math/verbal continuum described in the Marsh/Shavelson model. Also, measurement invariance tests of the ASCs showed that the ASC structures were invariant across LOI and provided validity for comparing I/E models across LOI groups.

three subjects were examined. The traditional two-subject I/E models were supported in the present investigation. When posited in the I/E mode including all three school subjects (Model TG4), Chinese and English had only very weak frame of reference effects on each other.

- 4. Based on the finding of (2) and (3), a higher order SEM model (Model TG5) with a second-order verbal factor was designed. In this model, the first-order factors of English and Chinese ASC were incorporated as indicators of a second-order verbal factor. The expected frame of reference effect was observed between the verbal and math domains. While there is room for disagreement as to which of these models best represents the data, it is important to emphasize

- 2-3. To investigate the effect of LOI on I/E model relationships, both the traditional two-subject models and models with all

that they are both consistent with the Marsh/Shavelson (Marsh & Shavelson, 1985) hypothesis and with the I/E models' prediction regarding the two general ASC factors: math ASC and verbal ASC. Clearly, English and Chinese were shown to be more similar to each other (consistent with the Marsh/Shavelson model), while contrasting with math.

5. Further to the above findings, multiple group analysis showed that the I/E model was fully equivalent in both English-LOI and Chinese-LOI schools, leading to the conclusion that the frame of reference effects applied similarly to both school types. It seems then that in Hong Kong, the English-LOI method does not moderate the effect of academic achievement on ASC.

In the present study, the correlations between math and verbal ASCs were moderately positive (math and English, 0.19; math and Chinese, 0.24; Table S2). They were slightly larger than those reported by Marsh, Kong, and Hau (2001), based on a different instrument and older students, but were still in line with correlations between math and verbal ASC in the meta-analysis (Möller, Pohlmann, et al., 2009) based on 69 studies (e.g., 0.27 for students up to Year 7, 0.01 for students in Years 7 to 9, and 0.09 for students beyond Year 9).

In the higher order I/E model, the negative cross-domain path coefficients leading from achievements to ASCs were higher than those observed in the first-order I/E model. The higher order ASC factor represents a common factor based on English and Chinese first-order ASC factors. Since the reliability issues associated with measurement errors at the item level were accounted for, the residual variances of first-order factors are indicative of any lack of agreement between English and Chinese ASCs. The increased regression weights associated with the higher order constructs represent a closer association between higher order ASC constructs and the corresponding higher order achievement factors. This is in agreement with the notion that a higher order verbal ASC may be a better representation for the close relationship between ASCs in Chinese and English. To our knowledge, this study is the first to examine an I/E model that is based on a higher order verbal ASC incorporating native and nonnative languages. In order to replicate this finding, future studies could be designed where higher order ASC and achievement measures are posited.

Although a higher order verbal ASC provided a good representation of the relation between English and Chinese ASCs, this does not imply that the English and Chinese ASCs were indistinguishable. If this were the case, a single factor for the Chinese and English ASCs would be sufficient to represent verbal ASC. In a supplemental test of this model, we combined Chinese and English ASCs into a single first-order factor, but this model clearly did not fit the data ($CFI = 0.83$, $TLI = 0.72$, $RMSEA = 0.153$), while the corresponding model with English and Chinese ASCs as separate factors fitted the data well (Models TGS1 and TGS2, supplemental material). The correlation between the two ASCs ($r = .45$ in Model TGS1, Table S2, supplemental material) was substantially less than 1.0, demonstrating that students clearly differentiated between native and foreign language ASCs. The higher order verbal ASC explained the substantial correlation between English

and Chinese ASC and was consistent with a multidimensional hierarchical representation of ASC.

In relation to our extension of the I/E model, the most intriguing finding is the statistically nonsignificant paths from English/Chinese achievement to Chinese/English ASC. This implies that the internal frame of reference does not apply to English and Chinese subjects. If a student performs well in English, this student's Chinese ASC will not be affected adversely. Similarly for a student excelling in Chinese, English ASC will not suffer, as would be the case for very distinct domains such as math and English/Chinese. This led to the development of the higher order I/E model (Model TG5), where English and Chinese ASCs were combined into a higher order ASC. The higher order I/E model supported the typical two domain (math and verbal) I/E model. The contrast effect was confirmed by the dimensional comparison process shown by the negative cross paths from math and verbal cross-domain achievement to ASC. The present study's results are in agreement with findings from the German study (Möller, Streblo, Pohlmann, & Köller, 2006). These authors found almost no effect from English (German) achievement to German (English) ASC, even though an earlier Hong Kong study (Marsh et al., 2001) reported small negative cross paths between English and Chinese at the beginning of high school. Nevertheless, based on our findings, the unique contribution of the present investigation was to combine a native language subject with a foreign language subject; this extension demonstrated support for the originally proposed I/E model theory.

Regarding the role of LOI, the present study showed that (a) ASC measurement properties were found to be invariant across LOIs; (b) in terms of the generalizability of the I/E model, no distinct differences were found across English- and Chinese-LOI schools. These findings have important implications for the development of the students' ASCs. The measurement invariance properties of the ASCs indicate that students from English- or Chinese-LOI schools perceived ASCs similarly and support the validity for the comparison of I/E models across LOI groups. The English-LOI students apparently shared a similar frame of reference process, both in terms of the external, social comparison process and the internal, dimensional comparison process. However, previous studies have shown that LOI might still have an effect on students' achievement and ASC, above and beyond the effect of the individual student's achievement.

It has been demonstrated that an LOI as foreign language is not necessarily effective for late immersion programs (Thomas & Collier, 1997; also see Willig, 1985). This is the case with Hong Kong, where the LOI changes in secondary education, rather than the more favorable form of early immersion, which starts in primary education (e.g., Marsh et al., 2000; Marsh, Hau, & Kong, 2002; Tam, 1980; Tung, 1990). The rationale for the learning of English, and for its use as the LOI, has been based on perceived pragmatic utilities, such as pursuing higher education and working as professionals in the business world and service sectors where English is extensively utilized (Evans, 2009; Li, 2002). Alongside Hong Kong's economic development as a world center for trade, English as an international business language has become increasingly important. However, even though the dual language educational system has a long history, the percentage of people who routinely use English in everyday activities is not substantial in Hong Kong. Apart from the fields of trade, administration, and

legislature and among the judiciary, English is not extensively used in everyday social interaction by the Chinese-speaking populace.

In the Hong Kong context, one of the drawbacks of teaching and learning through English is the difficulties students face in understanding and carrying out class tasks in a foreign language if their English skills are not already at a competent level. Whether bilingualism is beneficial or not therefore depends substantially on how competent the students are in the language of instruction. Concomitantly, it is self-evidently much easier for students to understand course content in their native language, and this better enables them to retain their learning motivation and preserve the quality of their teaching and learning activity. The same is true for teachers. Since English is also a second language to the teachers (Llewellyn, Hancock, Kirst, & Roeloffs, 1982), teachers in English-LOI classrooms tend to rely more on didactic approaches to teaching and are less effective in communicating abstract concepts to the students (Yip, Coyle, & Tsang, 2007). Indeed, students in nonnative-LOI classrooms tend to have lower engagement in classroom activities, to use fewer learning strategies (Salili & Lai, 2003), and to experience slower development in non-language-related outcomes (Marsh et al., 2000, 2002; see also Halle, Hair, Wandner, McNamara, & Chien, 2012). These difficulties are disadvantageous to the development of the students' learning and motivations to learn. Intervention programs designed to enhance both teaching and learning should also target ASCs, since these are reciprocally linked with academic achievement (Marsh & Yeung, 1997).

Following the above aspects in nonnative language LOI, in order to better promote the students' learning and the motivation to learn, it is particularly important for the educators to have a full understanding of the I/E model and its implications on educational practice. Based on the premise that the ASC facilitates many future academic accomplishments, aspirations, and choices (Marsh, 1991, 2007), maintaining a positive ASC is clearly crucial in helping students to reach their academic potential, and this has been shown to generalize both in English and Chinese LOI students (Marsh et al., 2002). This, however, would also require teachers and parents to understand more deeply how such process might be best supported. Nevertheless, when asked about the perceived domain-specific academic competencies of the students, teachers (Marsh & Craven, 1997) and parents (Dai, 2002) indicated that their perceptions of students' competency in different domains were based primarily on external processes of evaluation and did not differentiate between domains nearly as much as students did themselves. In other words, it appears that a student who is good in math would also be perceived by parents and teachers as being good at verbal subjects, and vice versa. This indicates a difference in the views that the parents and teachers take in terms of the internal frame of reference aspect of students' ASCs posited in the I/E model theory. When students show reduced effort or interest in learning in their weak subjects, teachers and parents need to intervene in order to facilitate students' learning: both by enhancing their academic ability and by restoring their ASC in their weaker domains. Similarly, it would be helpful for curriculum developers to build strong connections between different school subjects, to promote the cultivation of students' confidence and interest in learning all their academic subjects, instead of just their best ones. Based on our finding that the I/E model is

generalizable to both English and Chinese LOI students, this practice would be applicable to the students under both LOIs.

Strengths, Weaknesses, and Direction for Further Research

The present study had several particular methodological strengths. For instance, standard errors of parameter estimates were corrected in relation to the complex data structure. It is acknowledged in the social sciences that ignoring the design effect of the data can lead to biased estimates of standard errors. In the present study, only individual-level parameters were considered, so the hierarchical nature of the data structure was not modeled explicitly. Instead, complex design modeling was used. In this method, the parameter estimates remained the same as in single-level modeling, but the standard errors of the parameter estimates were corrected for students being nested within schools and classes.

A potential limitation of the present study is that the while 15% of the students from the Chinese LOI group were from low-ability band, participants from the low-ability band were not present in the English LOI group. However, this sampling composition is in agreement with the actual Hong Kong secondary school population in terms of LOI. In Hong Kong, in order for a school to adopt English-LOI practice, the students' academic ability needs to reach certain level in order for the English-LOI practice to be effective. Nevertheless, studies that can replicate our results based on samples with completely balanced ability compositions will certainly provide further robustness to our findings. Another feature of the sample that is worth noting in our study is that the ASC measures were collected approximately 1 year after the students started their secondary school. It is possible that changes in the processes related to the I/E model could take place depending on the duration of the immersion. Longitudinal studies that investigate the effect of LOI at different time points after students are immersed in new LOI environments will shed light on the processes that the effect of LOI takes place.

An interesting finding in the present study is the magnitude of some of the regression coefficients observed from the first-order I/E model (Model TG4). The regression coefficient relating Chinese achievement to Chinese ASC was smaller than the corresponding value in the English and math domains. Since Chinese, as the native language of Hong Kong, is used pervasively outside of school, it is likely that Chinese ASC is more broadly based, while math and English ASC are based primarily on achievement tests and performance in classroom settings. Certainly, students use Chinese language more frequently and on more diverse occasions compared to the utility of English and math. The meta-analysis (Möller, Pohlmann, et al., 2009) also showed that verbal achievement in native language predicted verbal ASC to a lesser extent compared to predictions from math achievement to math ASC. A study (Marsh, 1990) of Australian students showed that different components of English (native language) ASCs were only moderately correlated, this indicates that the English ASC itself might be able to be further divided. Hence, further in-depth information might be revealed if more domain specific assessments were implemented in Chinese ASC instruments, alongside the corresponding specific Chinese achievement measures.

The support for a higher order verbal factor incorporating English and Chinese in the full I/E model implies that a possible general verbal factor is in line with predictions in the Marsh/Shavelson (Marsh & Shavelson, 1985) model, where general math and verbal ASC factors are posited. Whether this would hold true for other verbal domain subjects is beyond the scope of the present investigation. However, it would be interesting to test whether other domains classified as verbal-based subjects, as posited by Marsh and Shavelson (1985; Figure 1B), such as history or foreign languages other than English, belong to the same factor as English and Chinese. Future studies could explore this by collecting data in a multilingual context such as Singapore, where many languages coexist in society. Similarly it might be reasonable, for example, that math and physics combine to form a single frame of reference. Extending this logic, it is still not clear whether two contrasting frames of reference (math vs. verbal) would be able to incorporate all school subjects (e.g., history, geography, computer studies, but also art, physical education, industrial arts and home economics). Furthermore, the answer may depend on the level and type of education the students receive. For example, university students who major in science might be more likely to distinguish between physical and biological sciences in terms of ASC formation than students at the secondary and particularly at the primary school levels. Future studies based on responses from students at different levels of education would provide evidence to clarify this issue.

Another interesting direction that might be relevant to the present investigation is the competency and affective components that are implied in the ASC instrument. Items pertaining to affect assess the level of interests and enjoyment the students have for a particular subject, whereas items pertaining to competency ask questions about whether they are competent and get good marks in a subject. Even though the correlations between competency and affect can be as high as 0.75 (e.g., Marsh, Craven, & Debus, (1999), the relationship between different components of ASC might be differentially correlated with other variables. For example, in a recent study based on a sample of German students from third to sixth grades (Arens, Yeung, Craven, & Hasselhorn, 2011), the competency component of the ASC was found to be more closely related with the achievement than the affect component. Since LOI has been shown to affect learning motivations (Salili & Lai, 2003), it is important to examine the applicability of the I/E model to these more specific components of ASC. To achieve this, the study design would involve assessing both competency and affect from different LOI students, on multiple school subjects and using multiple indicators for both components.

To date, not much research has looked at the generalizability of the I/E model outside of the context of ASC. It is possible that the internal and external comparison processes generalize across different psychological constructs such as those posited in the Student Approaches to Learning instrument (SAL; Marsh, Hau, Artelt, Baumert, & Peschar, 2006). Generalizability of the I/E model will be relevant to the concept of domain specificity, that is, how closely the constructs are correlated across different subject domains. If the construct's correlations across district domains are as low as ASC, then this construct is of high domain specificity, and it is likely that the frame of reference effect is present in the formation of this construct. Conversely, if a construct is of low domain specificity then the I/E model may be less or only partly applicable to this construct. Studies that examine domain specific-

ity and the generalizability of the I/E models to a wider range of psychosocial variables will provide valuable contribution to the understanding of the development of the students' learning and motivation.

References

- Arens, A. K., Yeung, A. S., Craven, R. G., & Hasselhorn, M. (2011). The twofold multidimensionality of academic self-concept: Domain specificity and separation between competence and affect components. *Journal of Educational Psychology, 103*, 970–981. doi:10.1037/a0025047
- Bong, M. (1998). Tests of the internal/external frames of reference model with subject-specific academic self-efficacy and frame-specific academic self-concepts. *Journal of Educational Psychology, 90*, 102–110. doi:10.1037/0022-0663.90.1.102
- Byrne, B. M. (2001). *Structural equation modeling with AMOS*. Mahwah, NJ: Erlbaum.
- Chiu, M. S. (2008). Achievements and self-concepts in a comparison of math and science: Exploring the internal/external frame of reference model across 28 countries. *Educational Research and Evaluation, 14*, 235–254. doi:10.1080/13803610802048858
- Chiu, M. S. (2012). The internal/external frame of reference model, big-fish-little-pond effect, and combined model for mathematics and science. *Journal of Educational Psychology, 104*, 87–107.
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures: New approaches to missing data. *Psychological Methods, 6*, 330–351. doi:10.1037/1082-989X.6.4.330
- Dai, D. Y. (2002). Incorporating parent perceptions: A replication and extension study of the internal-external frame of reference model of self-concept development. *Journal of Adolescent Research, 17*, 617–645. doi:10.1177/074355802237467
- Evans, S. (2009). The medium of instruction in Hong Kong revisited: Policy and practice in the reformed Chinese and English streams. *Research Papers in Education, 24*, 287–309. doi:10.1080/02671520802172461
- Evans, S. (2011). Historical and comparative perspectives on the medium of instruction in Hong Kong. *Language Policy, 10*, 19–36. doi:10.1007/s10993-011-9193-8
- Halle, T., Hair, E., Wandner, L., McNamara, M., & Chien, N. (2012). Predictors and outcomes of early versus later English language proficiency among English language learners. *Early Childhood Research Quarterly, 27*, 1–20. doi:10.1016/j.ecresq.2011.07.004
- Köller, O., Klemmert, H., Möller, J., & Baumert, J. (1999). Eine längsschnittliche Überprüfung des Modells des internal/external frame of reference [A longitudinal test of the internal/external frame of reference model]. *Zeitschrift für Pädagogische Psychologie, 13*, 128–134. doi:10.1024/1010-0652.13.3.128
- Li, D. (2002). Hong Kong parents' preference for English-medium education: Passive victims of imperialism or active agents of pragmatism. In A. Kirkpatrick (Ed.), *Englishes in Asia: Communication, identity, power and education* (pp. 29–62). Melbourne, Australia: Language Australia.
- Llewellyn, J., Hancock, G., Kirst, M., & Roeloffs, K. (1982). *A perspective on education in Hong Kong: Report by a visiting panel*. Hong Kong: Government Printer.
- Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*, 129–149.
- Marsh, H. W. (1990). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology, 82*, 623–636. doi:10.1037/0022-0663.82.4.623
- Marsh, H. W. (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels. *American Educational Research Journal, 28*, 445–480.

- Marsh, H. W. (1992). Content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology, 84*, 35–42. doi:10.1037/0022-0663.84.1.35
- Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self concept in educational psychology*. London, England: British Psychological Society.
- Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315–353). Hillsdale, NJ: Erlbaum.
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391–410. doi:10.1037/0033-2909.103.3.391
- Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology, 80*, 366–380. doi:10.1037/0022-0663.80.3.366
- Marsh, H. W., & Craven, R. G. (1997). Academic self-concept: Beyond the dustbowl. In G. D. P. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 131–198). San Diego, CA: Academic Press.
- Marsh, H. W., Craven, R., & Debus, R. (1999). Separation of competency and affect components of multiple dimensions of academic self-concept: A developmental perspective. *Merrill-Palmer Quarterly: Journal of Developmental Psychology, 45*, 567–601.
- Marsh, H. W., & Hau, K. T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64*, 364–390.
- Marsh, H. W., & Hau, K. T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Educational Psychology, 96*, 56–67. doi:10.1037/0022-0663.96.1.56
- Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology—Most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*, 311–360. doi:10.1207/s15327574ijt0604_1
- Marsh, H. W., Hau, K. T., & Kong, C. K. (2000). Late immersion and language of instruction in Hong Kong high schools: Achievement growth in language and non-language subjects. *Harvard Educational Review, 70*, 302–346.
- Marsh, H. W., Hau, K. T., & Kong, C. K. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. *American Educational Research Journal, 39*, 727–763. doi:10.3102/00028312039003727
- Marsh, H. W., & Köller, O. (2004). Unification of theoretical models of academic self-concept/achievement relations: Reunification of East and West German school systems after the fall of the Berlin Wall. *Contemporary Educational Psychology, 29*, 264–282. doi:10.1016/S0361-476X(03)00034-1
- Marsh, H. W., Kong, C. K., & Hau, K. T. (2001). Extension of the internal/external frame of reference model of self-concept formation: Importance of native and nonnative languages for Chinese students. *Journal of Educational Psychology, 93*, 543–553. doi:10.1037/0022-0663.93.3.543
- Marsh, H. W., Martin, A. J., & Hau, K. T. (2006). A multimethod perspective on self-concept research in educational psychology: A construct validity approach. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 441–456). Washington, DC: American Psychological Association. doi:10.1037/11383-030
- Marsh, H. W., & Shavelson, R. J. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*, 107–123. doi:10.1207/s15326985ep2003_1
- Marsh, H. W., & Yeung, A. S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology, 89*, 41–54. doi:10.1037/0022-0663.89.1.41
- Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35*, 705–738.
- Marsh, H. W., & Yeung, A. S. (2001). An extension of the internal/external frame of reference model: A response to Bong (1998). *Multivariate Behavioral Research, 36*, 389–420. doi:10.1207/S15327906389-420
- Möller, J., & Köller, O. (2001a). Dimensional comparisons: An experimental approach to the internal/external frame of reference model. *Journal of Educational Psychology, 93*, 826–835. doi:10.1037/0022-0663.93.4.826
- Möller, J., & Köller, O. (2001b). Frame of reference effects following the announcement of exam results. *Contemporary Educational Psychology, 26*, 277–287. doi:10.1006/ceps.2000.1055
- Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*, 1129–1167. doi:10.3102/0034654309337522
- Möller, J., Retelsdorf, J., Köller, O., & Marsh, H. W. (2011). The reciprocal internal/external frame of reference model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48*, 1315–1346. doi:10.3102/0002831211419649
- Möller, J., Streblow, L., & Pohlmann, B. (2006). The belief in a negative interdependence of math and verbal abilities as determinant of academic self-concepts. *British Journal of Educational Psychology, 76*, 57–70. doi:10.1348/000709905X37451
- Möller, J., Streblow, L., & Pohlmann, B. (2009). Achievement and self-concept of students with learning disabilities. *Social Psychology of Education, 12*, 113–122. doi:10.1007/s11218-008-9065-z
- Möller, J., Streblow, L., Pohlmann, B., & Köller, O. (2006). An extension to the internal/external frame of reference model to two verbal and numerical domains. *European Journal of Psychology of Education, 21*, 467–487. doi:10.1007/BF03173515
- Mui, F. L. L., Yeung, A. S., Low, R., & Jin, P. (2000). Academic self-concept of talented students: Factor structure and applicability of the internal/external frame of reference model. *Journal for the Education of the Gifted, 23*, 343–367.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological methodology, 25*, 267–316. doi:10.2307/271070
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Plucker, J. A., & Stocking, V. B. (2001). Looking outside and inside: Self-concept development of gifted adolescents. *Exceptional Children, 67*, 534–548.
- Pohlmann, B., & Möller, J. (2009). On the benefit of dimensional comparisons. *Journal of Educational Psychology, 101*, 248–258. doi:10.1037/a0013151
- Salili, F., & Lai, M. K. (2003). Learning and motivation of Chinese students in Hong Kong: A longitudinal study of contextual influences on students' achievement orientation and performance. *Psychology in the Schools, 40*, 51–70. doi:10.1002/pits.10069
- Schafer, J. L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]. University Park: Pennsylvania State University, Department of Statistics.

- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. doi:10.1037/1082-989X.7.2.147
- Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, 46, 407–441.
- Stapleton, L. M. (2006). Using multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 345–383). Greenwich, CT: Information Age.
- Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.
- Tam, P. T. K. (1980). A survey of the language mode used in teaching junior forms in Anglo-Chinese secondary schools in Hong Kong. *RELC Journal*, 11, 43–56. doi:10.1177/003368828001100104
- Tay, M. P., Licht, B. G., & Tate, R. L. (1995). The internal/external frame of reference in adolescents' math and verbal self-concepts: A generalization study. *Contemporary Educational Psychology*, 20, 392–402. doi:10.1006/ceps.1995.1026
- Thomas, W. P., & Collier, V. (1997). *School effectiveness for language minority students*. Washington, DC: National Clearinghouse for Bilingual Education.
- Tung, P. (1990). Why changing the medium of instruction in Hong Kong could be difficult. *Journal of Multilingual and Multicultural Development*, 11, 523–534. doi:10.1080/01434632.1990.9994436
- Willig, A. C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research*, 55, 269–317.
- Xu, M. (2010). *Frame of reference effects in academic self-concept: An examination of the big-fish-little-pond effect and the internal/external frame of reference model for Hong Kong adolescents* (Unpublished doctoral dissertation). University of Oxford, Oxford, England.
- Yeung, A. S., & Lau, I. C. (1998). *The internal/external frame of reference in the self-concept development of higher education students*. Paper presented at the Conference of the Higher Education Research and Development Society of Australasia, Auckland, New Zealand.
- Yeung, A. S., Lee, J. C., & Wong, H. (2001, April). *Testing Marsh's (1986) frame of reference model of self-concept with bilingual students*. Paper presented at the Annual meeting of the American Educational Research Association, New Orleans, LA.
- Yip, D. Y., Coyle, D., & Tsang, W. K. (2007). Evaluation of the effects of the medium of instruction on science learning of Hong Kong secondary students: Instructional activities in science lessons. *Education Journal*, 35, 77–107.

Received June 15, 2011

Revision received November 29, 2012

Accepted November 29, 2012 ■

The Role of Goal Attainment Expectancies in Achievement Goal Pursuit

Corwin Senko

State University of New York at New Paltz

Chris S. Hulleman

James Madison University

The current studies introduce the goal attainment expectancy construct to achievement goal theory. Three studies, 2 in college classrooms and the other using a novel math task in the laboratory, converged on the same finding. For mastery-approach goals and performance-approach goals alike, the harder the goal appeared to attain, the less likely participants were to pursue it, ultimately with negative repercussions to participants' task interest and achievement. Additionally, in each study, mastery-approach goals were generally considered easier to achieve than performance-approach goals. Study 3 also demonstrates that these judgments are highly amenable to cues in the situation: Mastery-approach goal expectancies are colored by the apparent complexity of the material to be learned. Multiple theoretical implications are considered, particularly for work on achievement goal antecedents and goal revision.

Keywords: achievement goals, expectancies, goal revision

Achievement goal theory, like most motivation theories, gives competence beliefs a pivotal role. The exact role is open to debate, however. Theorists initially treated competence beliefs as a moderating variable, positing that any achievement goal effects should depend upon the individual's self-perceived competence (Dweck, 1986; Nicholls, 1984). Elliot (1999, 2005) later conceptualized competence beliefs as an antecedent to goal pursuit, positing that students' general expectancy for doing well should dictate whether they pursue an achievement goal. Herein, we offer a friendly amendment to Elliot's model by splitting the general competence expectancy construct into specific and separate expectancies for attaining mastery goals or performance goals. In so doing, we identify several ways in which studying goal attainment expectancies could enrich achievement goal theory.

Achievement Goal Theory Overview

Achievement goal theory features mastery goals and performance goals. Both goals concern the pursuit of competence and the assessment of one's own skill level, but in different ways. When pursuing performance goals, people focus on outperforming peers and define success versus failure with normative standards. When pursuing mastery goals, they instead focus on developing their skills and define success versus failure with task-based or self-referential standards. Theorists eventually added a goal valence dimension to the theory (Elliot, 1999; Pintrich, 2000), thus allow-

ing each goal to be framed in either an approach manner (i.e., a mastery-approach goal to learn, a performance-approach goal to outperform others) or an avoidant manner (i.e., a mastery-avoidance goal to avoid failing to learn, a performance-avoidance goal to avoid doing worse than others).

Extensive research attests to the impact these goals have on academic achievement (e.g., class grades) and interest in the course material, as well as various cognitive, affective, and behavioral processes that may aid or hinder these two important educational outcomes. On the whole, mastery-avoidance and performance-avoidance goals are linked to maladaptive outcomes, such as poor grades and low interest, anxiety, and confusion about how to study (see Baranik, Stanley, Bynum, & Lance, 2010; Moller & Elliot, 2006). The two approach goals instead are linked to positive or neutral outcomes. Mastery-approach goals seem especially adaptive. For example, students who pursue these goals find the course material more interesting, persist longer, seek help when confused, and self-regulate effectively (Harackiewicz, Barron, Tauer, & Elliot, 2002; Ryan & Pintrich, 1997; Wolters, 2004). Surprisingly, though, mastery-approach goals are generally unlinked with actual achievement in the class, according to a comprehensive meta-analysis of approximately 100 studies (Hulleman, Schrager, Bodmann, & Harackiewicz, 2010). Performance-approach goals, by contrast, do appear to reliably predict high achievement in the class (Hulleman et al., 2010), but they typically are unrelated to course interest, deep learning strategies, and several of the other adaptive correlates of mastery-approach goals (Harackiewicz et al., 2002; Moller & Elliot, 2006).¹ Thus, mastery goals and performance goals, when framed in an approach manner,

This article was published Online First February 18, 2013.

Corwin Senko, Department of Psychology, State University of New York at New Paltz; Chris Hulleman, Department of Graduate Psychology, James Madison University.

Chris Hulleman is now at the Center for Advanced Study of Teaching and Learning, University of Virginia.

Correspondence concerning this article should be addressed to Corwin Senko, Department of Psychology, State University of New York at New Paltz, 600 Hawk Drive, New Paltz, NY 12561. E-mail: senkoc@newpaltz.edu

¹ Hulleman et al.'s (2010) meta-analysis revealed that goal relationships with grades depend on how the goals are measured. Mastery-approach goals predict high achievement when the goal measure includes challenge-seeking or interest elements but not when stripped of those confounding elements and defined strictly in terms of learning or task mastery. Performance-approach goals predict high achievement when the goal measure is framed in terms of normative comparison, as in the present research, but not when framed in terms of trying to appear smart to others.

appear to promote two largely independent sets of educational benefits (Barron & Harackiewicz, 2001). These effects have appeared in introductory and advanced courses, with different types of test formats, and in different countries (see Senko, Hulleman, & Harackiewicz, 2011). Consequently, some theorists advocate a “multiple goals” framework that considers pursuit of both goals more beneficial than pursuit of only one or the other, at least for some students (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002).² Throughout this article, we too focus on mastery-approach and performance-approach goals, though we also consider the two avoidance goals later in the article.

The benefits of these two approach goals compel theorists to identify antecedents that promote goal pursuit. The literature is rich with candidates that generally cluster into three types. One is individual differences among students, including, for example, their fear of failure and competitiveness (Baranik et al., 2010). The second is the classroom’s broader goal structure: A class can promote a mastery or performance structure based on the teacher’s methods of feedback and evaluation, how students are grouped, and the degree of autonomy given students (Ames, 1992), as well as the teacher’s practice of providing emotional support and encouraging social interaction among students (Patrick, Kaplan, & Ryan, 2011).³ Extensive research has explored how these two antecedents impact students’ personal achievement goals (for reviews, see Baranik et al., 2010; Moller & Elliot, 2006; Payne, Youngcourt, & Beaubien, 2007). The third type of antecedent is students’ competence beliefs (Baranik et al., 2010; Moller & Elliot, 2006). Having received far less research attention, it is the focus of this article.

Competence Beliefs and Achievement Goal Pursuit

Theorists initially designated competence beliefs a moderator of goal effects on achievement and other outcomes (Dweck, 1986; Nicholls, 1984). Mastery-approach goals and performance-approach goals, they hypothesized, should both yield benefits when students possess high self-perceived competence (i.e., a belief about one’s current ability level in that domain). But they should yield diverging effects when students lack self-perceived competence: Students pursuing mastery goals, owing largely to the malleable theories of intelligence undergirding these goals, were assumed to persevere by increasing effort, improving strategies, and so forth; by contrast, students pursuing performance goals, owing to their fixed theories of intelligence, were assumed to develop a more helpless response.⁴ Some studies supported this moderator hypothesis (e.g., Covington & Omelich, 1984; Elliott & Dweck, 1988), but others did not (e.g., Kaplan & Midgley, 1997; Miller, Behrens, & Greene, 1993).

This inconsistency in findings led Elliot (Elliot, 1999; Elliot & Church, 1997) to propose an alternative model for competence beliefs. He substituted general competence expectancy (i.e., a belief that one will do well in the class) for self-perceived competence and positioned it as an *antecedent* to goal pursuit. Whereas perceived competence and similar constructs (e.g., academic self-concept; Marsh, 1990) capture beliefs about one’s current capability, the competence expectancy construct, as a prospective judgment, presumably captures the combined beliefs about one’s current capability, the task’s difficulty, and the availability of contextual support to enable task success (e.g., effective teachers

whose methods support pursuit of the personal goal). In principle, then, the competence expectancy construct should be shaped by self-perceived competence but also extend beyond it, making it a more proximal and potent influence on students’ task-related behavior (Ferla, Valcke, & Cai, 2009).⁵

A few studies support Elliot’s model, each showing that high competence expectancy at the beginning of the semester predicts subsequent pursuit of mastery-approach goals and performance-approach goals and rejection of mastery-avoidance and performance-avoidance goals (e.g., Elliot & Church, 1997; Greene, Miller, Crowson, Duke, & Akey, 2004; Harackiewicz, Barron, Carter, Lehto, & Elliot, 1997; Miller et al., 1993). Yet most studies on goal antecedents have ignored the competence expectancy construct. Why has it gained so little traction? We believe the answer lies in its correlations with the mastery-approach and performance-approach goals: They tend to be equal in magnitude. Thus, this competence expectancy construct can successfully distinguish approach from avoidance goal framings (i.e., goal valence) but not mastery from performance goals (i.e., goal content). For a field preoccupied with mastery versus performance goal comparisons, this amounts to a sizable shortcoming in predictive value.

We see two related reasons for this limited predictive value. One is that the competence expectancy construct is quite broad, as illustrated by its measure: “I expect to do well in this class” and “I believe I will receive an excellent grade in this class” (Elliot & Church, 1997). The competence standard conveyed here is vague, not goal-specific. Moreover, people are likely to interpret this vague standard based on the goal in mind. Students focused on task mastery probably frame “doing well” in terms of learning, but those focused on outperforming others probably frame it in terms of normative ability. Further compounding this interpretation issue, mastery and performance goals may differ in perceived difficulty. Dweck and Elliott (1983), early in achievement goal the-

² Multiple goal effects can be revealed in several ways (Barron & Harackiewicz, 2001). The most obvious of them is an interaction effect between two goals, but those are uncommon (Harackiewicz, Barron, Pintrich, et al., 2002), and indeed we found none in our present studies. Multiple goal effects are more commonly shown through separate main effects of each goal on either (a) the *same* educational outcome (i.e., an additive model) or, especially, (b) *different* educational outcomes, such as interest versus achievement (i.e., a specialized model).

³ Goal structure was originally introduced as an antecedent to personal goals. Subsequent research showed that it can also affect student engagement directly, independent of students’ personal goals, or moderate the effects of students’ goals on student engagement (Murayama & Elliot, 2009).

⁴ Perceived competence in Dweck’s model, competence expectancies in Elliot’s model, and goal attainment expectancies in our model are all conceptually distinct from students’ theory of intelligence (Dweck, 1986), which is more concerned with beliefs about the malleability of competence than with beliefs about one’s level of competence.

⁵ Competence perception constructs abound in the literature: for example, expectancies (Wigfield & Eccles, 2000), self-efficacy (Pajares, 1996), personal agency beliefs (Ford, 1992), academic self-concept (Marsh, 1990), and so forth. A thorough comparison of them is beyond the scope of this article. Interested readers are directed to Bong and Clark (1999), Pajares (1996), and Williams (2010). Suffice it to say these constructs often are highly intercorrelated, especially when measured at similar levels of specificity (Schunk, 1984), making them more alike than different. Indeed, they are sometimes used interchangeably in motivation research (e.g., Wigfield & Eccles, 2000).

ory's development, proposed that it may be easier to attain a mastery goal than a performance goal because the standards used for defining success versus failure are more flexible with a mastery goal. In support of this premise, Senko and Harackiewicz (2005a) later found that students' perceived mastery-approach goals were easier than performance-approach goals for a word puzzle activity. Similarly, Martin, Marsh, Debus, and Malmberg (2008) found that students more strongly endorse mastery-approach goal items that lack mention of challenge. Thus, the two goals' separate standards (i.e., interpersonal vs. intrapersonal) for defining success, as well as their potential differences in difficulty, severely limit the capability of the competence expectancy construct to foretell pursuit of one goal over another. This limitation might easily be remedied, however, by narrowing our focus onto a goal-specific construct: goal attainment expectancies.

Goal Attainment Expectancies

Goal attainment expectancies feature in most other goal-based theories (Austin & Vancouver, 1996; Hollenbeck & Klein, 1987; Locke & Latham, 2002). In each, they play a vital role in the goal initiation and maintenance processes: One's adoption of and continued pursuit of a goal is tethered to beliefs that the goal is attainable with the resources available. The same should be true of achievement goals. Students' pursuit of mastery-approach goals should be tethered to a belief that they can master the task, while their pursuit of performance-approach goals should be tethered to a belief that they can outperform peers.⁶ Formally incorporating this construct into the model of achievement goal antecedents could enrich theorizing in three ways.

The obvious benefit of the goal attainment expectancy construct is its predictive value. Competence perception measures have greater predictive value when matched in specificity with the outcome under investigation. For example, task-specific self-efficacy judgments predict performance more successfully than do either generalized self-efficacy judgments or broad beliefs about one's capability in that domain (Ferla et al., 2009; Meece, Wigfield, & Eccles, 1990). Likewise, goal-specific expectancies, being more precise than the general competence expectancy construct, should provide greater predictive value. Quite simply, although mastery-approach and performance-approach goal attainment expectancies are both partly yoked to more general beliefs about one's capability in that domain, and thus correlate with each other (Jagacinski, Kumar, Boe, Lam, & Miller, 2010), they are also, as elaborated later, shaped by different beliefs about the task environment. They should therefore have different effects: Mastery-approach goal expectancies should anchor mastery-approach goal pursuit, while performance-approach goal attainment expectancies should anchor performance-approach goal pursuit. A recent laboratory study demonstrated this (Jagacinski et al., 2010), and the three present studies also test this postulate within the classroom and the laboratory alike.

A second benefit is that the goal attainment expectancy construct can sharpen theorizing about achievement goal revision (Fryer & Elliot, 2007; Muis & Edwards, 2009). Various theorists have speculated that students who pursue performance-approach goals will, when suffering setbacks that reduce confidence, switch to performance-avoidance goals or disengage altogether, thus hampering their educational experience (Brophy, 2005; Kaplan &

Maehr, 2007; Nicholls, 1984). Goal attainment expectancies likely play the central role in goal revision, yet they have been ignored in the prior research (though see Kumar & Jagacinski, 2011; Senko & Harackiewicz, 2005b). Including them in goal revision models would allow clearer mapping of when students do in fact revise their performance-approach goals. Additionally, focusing on goal expectancies raises the intriguing possibility that students also adjust their mastery-approach goal pursuit after setbacks. That is, for performance-approach goals and mastery-approach goals alike, students should decrease pursuit of the goal if it seems potentially too hard to attain. Study 2 provides a preliminary test by examining if change in goal attainment expectancies predicts concurrent change in pursuit of the corresponding approach goal.

A third benefit is that the goal attainment expectancy construct introduces possible goal antecedents that have thus far been overlooked by achievement goal theory. As noted earlier, competence expectancy judgments, plus goal attainment expectancies, are shaped not only by the student's self-perceived capability (i.e., ability as well as personal resources, such as time and effort, available to apply toward the goal) but also by perceptions of the task's difficulty and beliefs about the availability of contextual support aiding goal pursuit (Austin & Vancouver, 1996). The last two factors suggest new potential antecedents to achievement goals.

In particular, performance-approach goals, due to their reliance on normative standards, should hinge not only on beliefs about one's own capabilities but also on beliefs about classmates' capabilities because they, too, influence performance-approach goal attainment expectancies. As a rule, the more talented one's rivals appear, the lower one's expectancy for attaining this goal and the less likely one is to pursue it. Judgments of classmates' talent levels should be largely irrelevant to mastery-approach goal pursuit, however. Mastery-approach goals, due to their reliance on self-referential or task-based standards, should hinge more on beliefs about the complexity or difficulty of the material to be learned and on the quality of the teacher (e.g., clarity or availability for help). In general, the more complex or challenging the task appears, the lower one's expectancy for attaining this goal and the less likely one is to pursue it. These task characteristics should be largely irrelevant to one's performance-approach goal expectancies and goal pursuit, however, insofar as such things are a constant that should affect all classmates equally. Study 3 provides an initial test of this possibility by comparing the effects of simple versus complex task perceptions on goal attainment expectancies.

Educational Outcomes

We sought to test these ideas within a broader context of replication. This would allow reasonable confidence in the generalizability of any new effects involving goal attainment expectancies. To that end, we also examined goal relationships with students' achievement and interest. Not only are they arguably two of the most important educational outcomes, but their relationships with goals have been fairly consistent, with performance-approach

⁶ Nicholls (1984) implied that goal attainment expectancy could have an important role in achievement goals when he proposed that only the most confident students would dare pursue a performance goal. He made no parallel proposition for mastery goals, however.

goals predicting achievement and mastery goals predicting interest (for a meta-analysis, see Hulleman et al., 2010). For achievement, Studies 1 and 2 tested goal relationships with students' classroom grades, while Study 3 tested goal relationships with performance on a novel math task.

Because we are treating interest as an outcome variable, we focused on situational interest rather than individual interest, as the former is more responsive to situational cues (Hidi & Renninger, 2006; Renninger & Hidi, 2011). Specifically, we examined the two forms of situational interest in the classroom (Linnenbrink-Garcia et al., 2010): enjoyment of the material ("catch" situational interest) and the personal utility in the material ("hold" situational interest). These two types of situational interest are often correlated with each other and with mastery-approach goal pursuit (Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008) and together are catalysts for the development of an enduring individual interest characterized by self-guided exploration of the domain (Renninger & Hidi, 2011). Studies 1 and 2 measured both types of situational interest, whereas Study 3, like other laboratory studies, measured only catch situational interest.

Study 1

Overview and Hypotheses

Study 1 tested the links between university students' expectancies for attaining mastery-approach and performance-approach goals, the degree to which they pursue these two goals, and their interest and academic achievement. It served three purposes. The first is to compare the average perceived difficulty of mastery-approach and performance-approach goals. In accord with prior research (Senko & Harackiewicz, 2005a), *Hypothesis 1* is that, overall, students will report higher goal attainment expectancies for the mastery-approach goal than for the performance-approach goal.

The study's second purpose is to test if goal attainment expectancies trigger pursuit of the corresponding goal. *Hypothesis 2* is that, after removing shared variance between the two goal attainment expectancies (i.e., due to their underlying relationship with general competence perceptions and method variance), high expectancies for attaining the mastery-approach goal will predict greater pursuit of this goal but will be unrelated to performance-approach goal pursuit, while high expectancies for attaining a performance-approach goal will predict greater pursuit of this goal but be unrelated to mastery-approach goal pursuit.

The study's third purpose is to chart the enduring impact of goal attainment expectancies on distal educational outcomes. In accord with prior research (see Hulleman et al., 2010), *Hypothesis 3* is that mastery-approach goals will predict high interest material (catch and hold), and *Hypothesis 4* is that performance-approach goals will predict high achievement.

Method

Participants. Participants were 182 students (115 female, 67 male; M age = 18.8 years), predominantly freshmen (82%) and either Caucasian (71%) or African American (24%). They

were in any of five different general psychology sections, each with approximately 100 enrolled students, at a large southeastern United States university. There were no significant differences between courses in student demographics, student year, or course grade. They participated for extra credit.

Procedure and measures. Students completed a single questionnaire packet outside of class time during the 12th week of a 16-week semester. The packet assessed, in order, goal attainment expectancies, goal pursuit, interest in the course, and current course grade. With the exception of course grade, all measures were self-reported with a Likert-type scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (7). Students were instructed to think specifically about their general psychology class when answering all questions. They were assured confidentiality of their responses.

Achievement goal expectancies and pursuit. The two-item measures of mastery-approach goal attainment expectancy ("I am confident I can master the material in this class"; "Mastering the course material is difficult" [reversed]; $\alpha = .70$) and performance-approach goal attainment expectancy ("I am confident I can do better than most other students in this class"; "Performing better than most other students in this class is difficult" [reversed]; $\alpha = .68$) were developed for this study. The three-item measures of mastery-approach goal pursuit ("My goal in this class is to completely master the course material"; "It is important to me to understand the course material as thoroughly as possible"; "In a class like this, I prefer course material that really challenges me so that I can learn new things"; $\alpha = .75$) and performance-approach goal pursuit ("My goal in this class is to get a better grade than most of the other students"; "It is important for me to do well compared to others in this class"; "I would really like to do better than others in this class"; $\alpha = .80$) were based on two previously validated measures (Elliot & McGregor, 2001; Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000).

Situational interest and grades. We assessed situational interest using two four-item measures: catch interest ("I enjoy coming to lecture"; "The lectures in this class are entertaining"; "I like this class because the instructor is enthusiastic about the subject"; "The course material is interesting"; $\alpha = .90$) and hold interest ("I think what we are learning in this course is important"; "What I am learning in this class is relevant to my life"; "I can apply what we are learning in this course to real life"; "I find the content in this course personally meaningful"; $\alpha = .88$).

Participants reported their current course grade, excluding extra credit, using the common F (0) to A (4) grade scale. With the following distribution, the average grade ($M = 2.39$) was a low "C" (F = 3.8%, D = 18.1%, C = 29.1%, B = 33.0%, A = 15.9%).

Results

Table 1 provides the means and zero-order correlations among all measures. In support of Hypothesis 1, a paired-samples t test showed that students' goal attainment expectancies were higher for the mastery-approach goal ($M = 4.50$) than the performance-approach goal ($M = 4.25$), $t(181) = 3.11$, $p < .01$, $d = 0.20$.

Table 1
Means and Zero-Order Correlations for Study 1 Measures

Variable	1	2	3	4	5	6	7	8
1. Mastery-approach goal attainment expectancy	—							
2. Mastery-approach goal pursuit	.16	—						
3. Performance-approach goal attainment expectancy	.64	.11	—					
4. Performance-approach goal pursuit	.14	.31	.37	—				
5. Catch interest	.10	.36	.13	.06	—			
6. Hold interest	.06	.47	.06	.12	.53	—		
7. Current course grade	.49	.01	.56	.25	.18	.09	—	
8. Gender	-.10	.17	-.07	.00	.07	.08	.06	—
<i>M</i>	4.50	4.73	4.25	4.93	5.27	4.92	2.39	
<i>SD</i>	1.25	1.16	1.30	1.14	1.46	1.08	1.08	
α	.70	.75	.68	.80	.90	.88	—	

Note. $N = 182$. For gender, 0 = male, and 1 = female. Correlations $> .14$ are significant ($p < .05$). α = internal reliability (Cronbach's alpha) provided for all self-report measures using multiple items.

We tested the remaining hypotheses in two sets of regression analyses.⁷ The first set examined the antecedents of goal pursuit, in particular if a higher expectancy for attaining a goal predicted greater pursuit of that goal (Hypothesis 2). The second examined the consequences of goal pursuit, in particular if the mastery-approach and performance-approach goals were associated with high interest and achievement, respectively (Hypotheses 3 & 4).

Data were obtained from five classes, too few to use hierarchical linear modeling techniques to examine nested effects (Raudenbush & Bryk, 2002). Following best practices for such cases, we therefore aggregated the data and accounted for potential instructor effects in all regression analyses by creating four dummy codes to represent the five instructors. One limitation to this approach, however, is that it cannot assess within-class dependence of observations; fortunately, such concerns should be reduced if the next two studies replicate the findings reported here. None of these dummy codes interacted significantly with the two goal attainment expectancy variables or with the two goals, and so those interaction terms were trimmed from the final model.

In all regression analyses in this article, unreported effects were nonsignificant ($p > .10$), and all indirect effects (B) are bootstrap estimates (based on 5,000 trials) with bias-corrected and accelerated 95% confidence intervals (CI; Preacher & Hayes, 2004).

Antecedents to goal pursuit. To test if goal attainment expectancies promote goal pursuit, we separately regressed mastery-approach goals and performance-approach goals onto a *Goal Antecedents Model* comprising the two goal attainment expectancy terms, plus, as covariates, the four instructor dummy codes and participants' gender.⁸

Mastery-approach goals. The overall model for mastery-approach goal pursuit was significant, $F(7, 174) = 2.81, p < .01, R^2 = .10$. Supporting Hypothesis 2, students were more likely to pursue a mastery-approach goal in their class if they felt relatively confident that they could attain it, $F(1, 174) = 4.03, p < .05, \beta = .19$. Importantly, performance-approach goal attainment expectancy was unrelated to mastery-approach goal pursuit. Additionally, women were more likely than men to pursue this goal for their class, $F(1, 174) = 5.19, p < .05, \beta = .17$, and one of the instructor dummy codes was also significant ($p < .05$), indicating that

students in this class were more likely than those in other classes to pursue mastery-approach goals.

Performance-approach goals. The overall model for performance-approach goal pursuit was significant, $F(7, 174) = 5.03, p < .001, R^2 = .17$. Further supporting Hypothesis 2, students were more likely to pursue a performance-approach goal in their class if they felt relatively confident that they could attain it, $F(1, 174) = 23.39, p < .001, \beta = .43$. Mastery-approach goal attainment expectancy was unrelated to performance-approach goal pursuit.

Consequences of goal pursuit. To test the consequences of goal pursuit, we regressed the interest measures (catch and hold) and grades onto a *Goal Consequences Model* comprising the two goals, plus the terms from the Goal Antecedents Model. Note that in this and the next two studies, the mastery by performance interaction was nonsignificant and was thus excluded from all regression models.

Catch interest. The overall model for catch interest was significant, $F(9, 172) = 13.83, p < .001, R^2 = .42$. Matching Hypothesis 3, students pursuing a mastery-approach goal reported high catch interest, $F(1, 172) = 16.36, p < .001, \beta = .26$. This goal effect also provided an indirect effect for mastery-approach goal attainment expectancy ($B = 0.08$), 95% CI [.02, .17]. Additionally, three of the instructor dummy codes were also significant

⁷ Given the medium sample size and our desire to explore all possible effects involving the two new measures, multiple regression is more appropriate than structural equation modeling (Kline, 2005). We finish, however, with a supplemental single structural equation model to test overall fit of the final model implied by regression findings (Elliot & Church, 1997).

⁸ Because of the relatively high correlation among the two expectancy measures ($r = .64$), we conducted multicollinearity tests in these regression analyses. By conventional standards, multicollinearity would be indicated by *tolerance statistics* under .20 and *variance inflation factor (VIF) statistics* above 3.0. Our tests showed little evidence of multicollinearity: Tolerance statistics for both of the expectancy variables were above .55, and the VIF statistics were below 1.95. This was also true of all analyses reported in Studies 2 and 3.

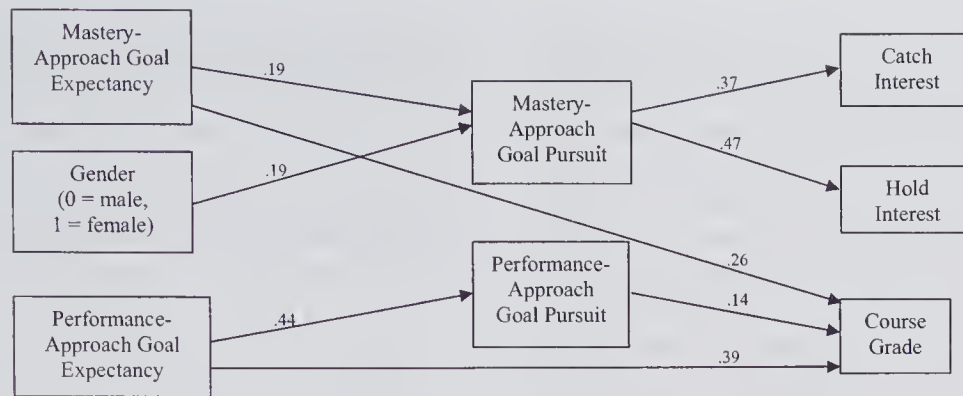


Figure 1. Full path model illustrating the goal antecedents and goal consequences pathways in Study 1. All paths are standardized regression coefficients ($p < .05$). For presentational clarity, nonsignificant paths, correlations, and instructor dummy codes are omitted.

($p < .05$), indicating that students in these classes reported higher catch interest than students in the other classes.

Hold interest. The model for hold interest was significant too, $F(9, 172) = 6.01$, $p < .001$, $R^2 = .24$. Students pursuing a mastery-approach goal found the material personally meaningful, $F(1, 172) = 39.48$, $p < .001$, $\beta = .46$, and this goal effect provided a significant indirect effect for mastery-approach goal attainment expectancy ($B = 0.11$), 95% CI [.02, .22].

Current grade. The model for course grade was significant, $F(9, 172) = 11.93$, $p < .001$, $R^2 = .38$. Matching Hypothesis 4, students pursuing a performance-approach goal earned high grades, $F(1, 172) = 3.62$, $p = .05$, $\beta = .14$. Additionally, mastery-approach goal attainment expectancy, $F(1, 172) = 10.38$, $p < .01$, $\beta = .26$, and performance-approach goal attainment expectancy, $F(1, 172) = 21.52$, $p < .001$, $\beta = .39$, also predicted high grades, the latter not only directly but also indirectly via its impact on performance-approach goal pursuit ($B = 0.11$), 95% CI [.02, .22]. Finally, women earned higher grades than men, $F(1, 172) = 4.48$, $p < .05$, $\beta = .13$.

Model fit. Figure 1 diagrams the results described above.⁹ Using AMOS 16.0 (Arbuckle, 2007), we tested the fit of this model with a structural equation model (SEM). Goal attainment expectancies were allowed to correlate, as were the disturbance terms for the two goals and for the two interest measures due to shared method variance (Kline, 2005).

According to conventional statistical criteria, a normed chi-square value (χ^2/df) below 3.0, a comparative fit index (CFI) value above .90, a root-mean-square error of approximation (RMSEA) under .08, and a standardized root-mean-square residual (SRMR) under .08 together indicate an adequate fit for a model (Kline, 2005). Our model met these criteria, thus showing it fit the data adequately ($\chi^2/df = 1.61$, CFI = .94, RMSEA = .06, SRMR = .07).

Summary

Each hypothesis was supported. Matching Hypothesis 1, participants, on average, held higher expectancies for attaining the mastery-approach goal than the performance-approach goal. Matching Hypothesis 2, goal attainment expectancies were antecedents to goal pursuit: The easier the mastery-approach goal or the performance-approach goal appeared to be to attain, the more likely students were to pursue those goals, respectively. Finally,

matching Hypotheses 3 and 4, the two goals were associated with distinct educational benefits: Mastery-approach goals predicted high situational interest in the course, while performance-approach goals predicted high course grades.

The two goal expectancies also predicted high grades. Though unexpected, this finding is sensible in light of the fact that these measures were collected after the midpoint of the semester, by which time students had established a performance pattern with which to anchor their expectancies for the class. Regardless, the more vital finding is that performance-approach goals predicted grades even when controlling for these expectancies, thus ruling out the possibility that this relationship merely reflects underlying confidence (Brophy, 2005).

The covariates had some effects as well. Women were more likely than men to pursue a mastery-approach goal, and women earned a higher grade, both of which replicate prior research (Harackiewicz et al., 1997). Additionally, students in one general psychology section were more likely than students in other sections to pursue a mastery-approach goal and to report high catch interest, thus testifying to the instructor's power to shape student motivation.

Study 2

The primary limitation of Study 1 is that it measured goal expectancies and goal pursuit at mid-semester, by which time students had already received some competence feedback in the course. Such late timing clouds the relationship that course grade has with expectancies and goal pursuit. For our purposes in this article, that relationship is much less important than the novel one between expectancies and goal pursuit. But the late timing of measures is nonetheless a notable shortcoming. Study 2 corrects this problem. It measured goal attainment expectancies and goal pursuit early in the semester, well before students obtain any competence feedback.

We also measured students' goal attainment expectancies and goal pursuit a second time, at mid-semester. This longitudinal design allowed us to extend prior work on goal stability. Students'

⁹ For simplicity, the figure omits effects due to instructor differences, but those effects were included in the model fit test. The same is true for Study 2.

goal pursuit early in the semester strongly predicts their pursuit of the same goal later in the semester, thus indicating relatively high stability in goal pursuit (Fryer & Elliot, 2007; Senko & Harackiewicz, 2005b). This is to be expected insofar as the typical classroom context (i.e., instructional style, peer composition, and evaluation method) probably remains fairly constant during the semester. Yet the sample-wide means for goal pursuit often decline during the semester. This suggests that, despite the overall pattern of stability, many students do revise their goals downward, presumably because of their ongoing experience (e.g., exam feedback; fatigue) in the class. Their goal attainment expectancies may be crucial to this revision. With Study 2, therefore, we attempt not only to replicate this pattern but also to examine if *changes* in goal attainment expectancies predict corresponding changes in goal pursuit.

Hypotheses

Hypothesis 1, retained from Study 1, is that students will report higher goal attainment expectancies for the mastery-approach goal than for the performance-approach goal. The next two hypotheses concern goal stability versus change. Replicating prior research, *Hypothesis 2* is that goal pursuit will generally be stable during the semester, and *Hypothesis 3* is that, despite this overall stability, pursuit of the two approach goals will decline during the semester.

The next two hypotheses concern the link between goal attainment expectancies and goal pursuit. Replicating Study 1, *Hypothesis 4* is that, after removing shared variance between the two goal attainment expectancies, high expectancies for attaining the mastery-approach goal pursuit will predict greater pursuit of this goal but will be unrelated to performance-approach goal pursuit, while expectancies for attaining a performance-approach goal will predict greater pursuit of this goal but be unrelated to mastery-approach goal pursuit. This should occur at Time 1 and Time 2 alike. Of course, as noted above (Hypothesis 3), some students adjust their goal pursuit during the semester, presumably because of experiences that alter their goal attainment expectancies. Accordingly, *Hypothesis 5* is that, for both goals, change in goal attainment expectancies will predict corresponding changes in goal pursuit.

Finally, consistent with Study 1, *Hypotheses 6* and *7* are that mastery-approach goals will predict high situational interest in the course material, and performance-approach goals will predict high grades. These effects should occur for the Time 1 and Time 2 goals alike, but when tested together, the Time 2 goals, due to their greater sensitivity to students' ongoing course experience, should prove more potent, perhaps even eclipsing the Time 1 goal effects.

Method

Participants. Participants were 230 students (165 female, 65 male; M age = 18.9 years) enrolled in four general psychology courses (300–400 students in each) at a large Midwestern United States university. They participated in return for extra credit. Participants were predominantly freshmen (54%) or sophomores (35%) and Caucasian (90%).

Procedure and measures. Students completed an online questionnaire outside of class time during the 2nd (Time 1) and 8th (Time 2) weeks of classes. The survey assessed, in order, achieve-

ment goal attainment expectancies and achievement goal pursuit. During the 14th week of the semester (Time 3) students completed the interest measures. All self-report measures used a Likert-type scale ranging from *Strongly Disagree* (1) to *Strongly Agree* (7). Students were instructed to think specifically about their general psychology class when answering all questions.

Achievement goal expectancies and pursuit. Participants read a performance-approach goal description:

Our research has found that students sometimes pursue the goal of doing better than other students in their class. We would like you to think about how the goal of *doing better than other students in Psychology 202* applies to you personally.

Afterward, they reported their expectancy for attaining this goal ("This goal will be hard for me to achieve in Psychology 202" [reversed]; "It will be easy to accomplish this goal in Psychology 202"; "It will be difficult to do better than others in Psychology 202" [reversed]; $\alpha_{\text{Time 1}} = .81$; $\alpha_{\text{Time 2}} = .86$). They then read a mastery-approach goal description:

Our research has found that students sometimes pursue another goal—the goal of understanding the course content as thoroughly as possible in their class. Now we would like you to think about how the goal of *learning as much as possible in Psychology 202* applies to you personally.

They then reported their expectancy for attaining this, using the same items as above ($\alpha_{\text{Time 1}} = .80$; $\alpha_{\text{Time 2}} = .81$).

They then completed a modified version of the Achievement Goal Questionnaire (Elliot & McGregor, 2001), with four-item measures of performance goal pursuit ("My goal in this class is to get a better grade than most of the other students"; "It is important for me to do well compared to other students in this class"; "I really don't care how I do compared to other students in this class" [reversed]; "I want to do better than other students in this class"; $\alpha_{\text{Time 1}} = .92$; $\alpha_{\text{Time 2}} = .93$) and mastery goal pursuit ("My goal in this class is to learn as much as I can"; "Completely mastering the material in this class is important to me"; "I want to learn as much as possible in this class"; "The most important thing for me is to understand the content in this class as thoroughly as possible"; $\alpha_{\text{Time 1}} = .92$; $\alpha_{\text{Time 2}} = .92$).

Situational interest. For situational interest, we used four-item measures of catch interest ($\alpha = .89$) and hold interest ($\alpha = .94$). They were identical to those from Study 1, except that we replaced a catch interest item in Study 1 ("I like this course because my instructor is enthusiastic") with a new item ("Lectures in this class drag on forever" [reversed]) to better match developments in the measurement of situational interest (Linnenbrink-Garcia et al., 2010).

Course grade. Students' final course grades were collected from the instructors, again on a 0–4 scale. With the following distribution, the average grade ($M = 3.02$) was a "B–" ($F = 0\%$, $D = 1.7\%$, $C = 21.3\%$, $C+/B– = 13.9\%$, $B = 19.1\%$, $B+/A– = 21.3\%$, $A = 22.6\%$).

Results

Table 2 provides the means and correlations among all measures. Matching Hypothesis 1, paired-samples t tests showed that students' goal attainment expectancies were higher for the

Table 2
Means and Zero-Order Correlations Among All Study 2 Measures

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. T1 mastery-approach goal attainment expectancy	—											
2. T1 mastery-approach goal pursuit	.22	—										
3. T2 mastery-approach goal attainment expectancy	.50	.14	—									
4. T2 mastery-approach goal pursuit	.12	.78	.19	—								
5. T1 performance-approach goal attainment expectancy	.52	.07	.32	.02	—							
6. T1 performance-approach goal pursuit	.03	.02	-.08	-.03	-.04	—						
7. T2 performance-approach goal attainment expectancy	.37	-.10	.47	-.06	.56	-.03	—					
8. T2 performance-approach goal pursuit	.05	.02	-.05	.06	.00	.72	.09	—				
9. Catch interest	.02	.29	.21	.39	.02	-.06	.14	.02	—			
10. Hold interest	.12	.43	.15	.50	.03	.01	.06	.12	.58	—		
11. Final course grade	-.03	.09	.16	.17	-.01	.10	.26	.21	.35	.31	—	
12. Gender	.01	.18	.00	.20	-.16	-.20	-.11	-.14	.09	.05	.05	—
<i>M</i>	3.97	5.49	3.99	5.22	3.59	5.71	3.54	5.51	4.68	5.37	3.02	
<i>SD</i>	1.22	1.07	1.18	1.23	1.15	1.13	1.23	1.19	1.45	1.24	0.77	
α	.81	.92	.81	.92	.81	.92	.86	.93	.89	.94	—	

Note. $N = 230$. For gender, 0 = male, and 1 = female. Correlations $> .12$ are significant ($p < .05$). α = internal reliability (Cronbach's alpha) provided for all self-report measures using multiple items; T = time.

mastery-approach goal ($M_{\text{Time 1}} = 3.97$; $M_{\text{Time 2}} = 3.99$) than the performance-approach goal ($M_{\text{Time 1}} = 3.59$; $M_{\text{Time 2}} = 3.54$) at both Time 1, $t(229) = 4.98$, $p < .01$ ($d = 0.32$), and Time 2, $t(229) = 5.50$, $p < .01$ ($d = 0.37$). In support of Hypothesis 2, the test-retest correlations for mastery-approach goal pursuit ($r = .78$) and performance-approach goal pursuit ($r = .72$) indicate that pursuit of each goal was generally stable. In support of Hypothesis 3, however, mean levels for mastery-approach pursuit and performance-approach goal pursuit declined significantly over time, $t(229) = 5.44$, $p < .001$ ($d = 0.24$), and $t(229) = 3.52$, $p < .01$ ($d = 0.17$), respectively, thus suggesting that some students' may have revised their goals due to their course experience.

Antecedents to Time 1 goal pursuit. To test whether goal attainment expectancies anchor Time 1 goal pursuit (Hypothesis 4), we separately regressed Time 1 mastery-approach and performance-approach goals onto a *Time 1 Goal Antecedents Model* that included both of the Time 1 goal attainment expectancy terms, plus, as covariates, participant's gender and three dummy codes representing the four instructors.

Time 1 mastery-approach goals. The model for mastery-approach goals was significant, $F(6, 223) = 4.71$, $p < .001$, $R^2 = .11$. Matching Hypothesis 4, students were more likely at Time 1 to pursue a mastery-approach goal if they held a high expectancy for attaining it, $F(1, 223) = 9.52$, $p < .01$, $\beta = .23$. Women were also more likely than men to pursue mastery-approach goals at Time 1, $F(1, 223) = 8.01$, $p < .01$, $\beta = .18$. One of the instructor dummy codes was also significant ($p < .05$), indicating that students in one class reported higher levels of mastery-approach goal pursuit than in other classes. Importantly, the performance-approach goal attainment expectancy was unrelated to mastery-approach goal pursuit.

Time 1 performance-approach goals. The model for performance-approach goals was significant, $F(6, 223) = 2.20$, $p < .05$, $R^2 = .06$. As in Study 1, men were more likely than women to pursue performance-approach goals at Time 1, $F(1, 223) = 10.86$, $p < .01$, $\beta = -.22$. In contrast to Hypothesis 4, however,

performance-approach goal expectancies failed to predict pursuit of this goal at Time 1 ($p > .10$).

Antecedents to Time 2 goal pursuit. To test whether goals are responsive to shifts in attainment expectancies, we hierarchically regressed the Time 2 mastery-approach and performance-approach goals onto a *Time 2 Goal Antecedents Model* that included two steps. Step 1 comprised all terms from the Time 1 Goal Antecedents Model plus both Time 1 goals. Step 2 comprised both Time 2 achievement goal attainment expectancy terms. Thus, Step 1 provides another test of overall stability in goal pursuit (Hypothesis 2), while Step 2, by controlling for the Time 1 goal attainment expectancy measures and goal measures, tests whether changes to goal attainment expectancies predict corresponding changes in goal pursuit (Hypothesis 5).¹⁰

Time 2 mastery-approach goals. The model for Time 2 mastery-approach goals was significant, $F(8, 221) = 49.52$, $p < .001$, $R^2 = .64$, with Step 2 explaining a significant increase in variance ($\Delta R^2 = .02$, $p < .05$). Step 1 showed that one of the instructor dummy codes was significant ($p < .001$), indicating that students in one class reported higher levels of mastery-approach goal pursuit than did students in other classes. More importantly, further supporting Hypothesis 2, Time 1 mastery-approach goals strongly predicted Time 2 mastery-approach goals, thus indicating goal stability over time, $F(1, 219) = 305.05$, $p < .001$, $\beta = .75$. Moreover, matching Hypothesis 5, Step 2 showed that the Time 2 mastery-approach goal attainment expectancy was a significant predictor, indicating that students who decreased in their mastery-approach goal attainment expectancy from Time 1 to Time 2 also decreased their mastery-approach goal pursuit at Time 2, $F(1, 219) = 7.56$, $p < .01$, $\beta = .14$.

¹⁰ An alternative approach is to use measures of change (i.e., Time 2–Time 1) in goal attainment expectancies ($\alpha = .67$ for each) as predictor variables (e.g., Jagacinski et al., 2010). Supplemental analyses using that approach verified the findings presented here. We chose the current approach to simplify the model fit tests presented later.

Time 2 performance-approach goals. The model for Time 2 performance-approach goals was significant, $F(8, 221) = 30.50$, $p < .001$, $R^2 = .53$, with Step 2 explaining a significant increase in variance ($\Delta R^2 = .02$, $p < .05$). Further supporting Hypothesis 2, Step 1 showed that Time 1 performance-approach goals strongly predicted Time 2 performance-approach goals, demonstrating stability in goal pursuit over time, $F(1, 221) = 226.71$, $p < .001$, $\beta = .72$. Moreover, matching Hypothesis 5, Step 2 showed that the Time 2 performance-approach goal attainment expectancy was a significant predictor, indicating that students who decreased in their performance-approach goal attainment expectancy from Time 1 to Time 2 also decreased their performance-approach goal pursuit at Time 2, $F(1, 219) = 8.98$, $p < .01$, $\beta = .18$.

Consequences of goal pursuit. To test the consequences of goal pursuit, we regressed situational interest (catch and hold) and grades onto a *Goal Consequences Model* that comprised two steps. Step 1 included all terms from the Time 2 Goal Antecedents Model described above, while Step 2 included the Time 2 mastery-approach and performance-approach goals. This hierarchical approach allowed separate tests of the effects of the Time 1 and Time 2 goals. For both time points, we expected mastery-approach goals to predict high interest (Hypothesis 6) and performance-approach goals to predict high grades (Hypothesis 7). But the Time 2 goal effects should be stronger and perhaps eclipse the Time 1 goal effects because the Time 2 goals capture students' ongoing experience in the class, as demonstrated in the above analyses of goal change.

Catch interest. The model for catch interest was significant, $F(10, 219) = 7.68$, $p < .001$, $R^2 = .26$, with Step 2 explaining a significant increase in variance ($\Delta R^2 = .05$, $p < .001$). Significant effects at Step 1 for Time 2 mastery-approach attainment expectancies, $F(1, 219) = 5.18$, $p < .05$, $\beta = .17$, and Time 2 performance-approach attainment expectancies, $F(1, 219) = 5.88$, $p < .001$, $\beta = .19$, show that students reported lower catch interest if their expectancies for attaining either goal decreased over time. More important, matching Hypothesis 6, students who pursued mastery-approach goals at Time 1 found the course to be more enjoyable at the end of the semester, $F(1, 219) = 23.10$, $p < .001$, $\beta = .35$. Step 2, however, showed that this effect was completely removed ($\beta = .01$, $p > .10$) and mediated ($B = 0.45$, 95% CI [.22, .73]) by Time 2 mastery-approach goals, $F(1, 217) = 16.13$, $p < .001$, $\beta = .40$. One of the instructor dummy codes was also significant ($p < .05$), indicating that students in that class reported lower levels of catch interest than did students in the other classes.

Hold interest. The results for hold interest parallel those for catch interest. The overall model was significant, $F(10, 219) = 8.93$, $p < .001$, $R^2 = .29$, with Step 2 explaining a significant increase in variance ($\Delta R^2 = .06$, $p < .001$). As expected, Step 1 showed that students who pursued mastery-approach goals at Time 1 found the course to be more personally meaningful at the end of the semester, $F(1, 219) = 48.77$, $p < .001$, $\beta = .43$. Step 2, however, showed that this effect was removed ($\beta = .17$, $p = .07$) and mediated ($B = 0.34$, 95% CI [.17, .56]) by the Time 2 mastery-approach goal, itself a strong predictor of hold interest, $F(1, 217) = 13.22$, $p < .01$, $\beta = .35$. One of the instructor dummy codes was also significant ($p < .05$), indicating that students in that class found the course to be less personally meaningful than did students in the other classes.

Final course grade. The overall model for course grade was significant, $F(10, 219) = 4.35$, $p < .001$, $R^2 = .17$, with Step 2 explaining a significant increase in variance ($\Delta R^2 = .04$, $p < .01$). A significant Time 2 performance-approach goal attainment expectancy effect indicates that students earned higher grades if their expectancy for attaining this goal increased over time, $F(1, 219) = 22.62$, $p < .001$, $\beta = .36$. Independent of this, and consistent with Hypothesis 7, students who pursued performance-approach goals at Time 1 earned significantly higher course grades, $F(1, 219) = 3.84$, $p = .05$, $\beta = .11$. Step 2, however, showed that this effect was removed ($\beta = -.01$, $p > .10$) and mediated ($B = 0.10$, 95% CI [.01, .19]) by the Time 2 performance-approach goal, itself a significant predictor of grades, $F(1, 217) = 4.86$, $p < .05$, $\beta = .20$.

Model fit. A supplemental SEM tested the overall fit of the findings described above (see Figure 2). The Time 1 goal attainment expectancies were allowed to correlate and to predict their Time 2 versions, and the disturbance terms were allowed to correlate for the Time 2 expectancies, for the two goals at both time points, and for the two interest measures. This model fit the data adequately ($\chi^2/df = 2.59$, CFI = .90, RMSEA = .08, SRMR = .08).

Summary

Most of the Study 2 hypotheses were supported. First, matching Hypothesis 1, at Time 1 and Time 2 alike, students held higher expectancies for attaining the mastery-approach goal than the performance-approach goal. Second, although pursuit of each goal remained relatively stable over time, thus matching Hypothesis 2, each also declined on average for the sample, thus matching Hypothesis 3 as well. Of greater importance, goal attainment expectancies were again an antecedent to goal pursuit. Matching Hypothesis 4, the easier the mastery-approach goal initially appeared (Time 1), the more likely students were to pursue it; however, to our surprise, the expected parallel pattern was not observed for the performance-approach goal. Nevertheless, when examining shifts in expectancies over time, we found that students' maintenance or revision of their achievement goals at mid-semester (Time 2) hinged on their ongoing goal attainment expectancies. In particular, matching Hypothesis 5, reductions in mastery-approach goal attainment expectancies or in performance-approach goal attainment expectancies predicted corresponding reductions in pursuit of the respective goal. Finally, matching Hypotheses 6 and 7, mastery-approach goals and performance-approach goals predicted high situational interest and grades, respectively, thus demonstrating the distal impact of goal attainment expectancies.

Several gender and instructor effects emerged too. Women were more likely than men to pursue mastery-approach goals, as in Study 1 and prior research (Harackiewicz et al., 1997), and also less likely to pursue performance-approach goals. Additionally, as further evidence of teachers' impact on student motivation, one course's students more strongly pursued mastery-approach goals, while another course's students reported higher catch and hold interest.

Study 3

Study 3 replicates and extends the first two studies in two important ways. First, it used a controlled laboratory setting to test

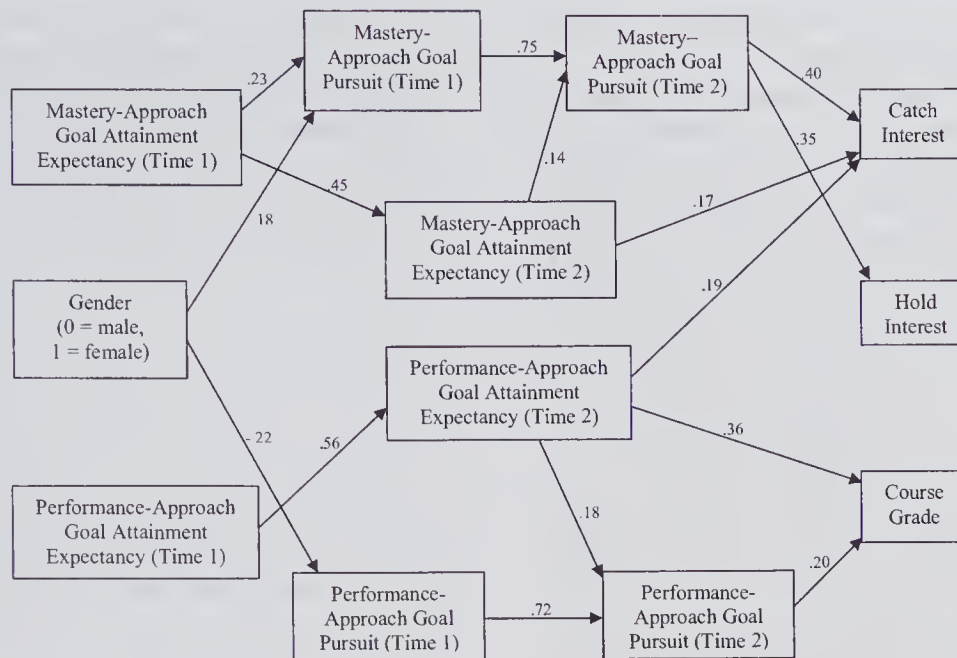


Figure 2. Path model illustrating the goal antecedents and goal consequences pathways in Study 2. All paths are standardized regression coefficients ($p < .05$). For presentational clarity, nonsignificant paths, within-time correlations, and instructor dummy codes are omitted.

the hypothesized effects. Second, and more important, it tested a new proposition suggested by the classroom findings. If goal attainment expectancies influence achievement goal pursuit, then it is vital to consider features of the learning situation that may affect goal attainment expectancies in the first place. For example, the perceived complexity of a new task to be learned might affect students' expectancies for attaining a mastery-approach goal, which defines success with task-based standards, but it should be largely irrelevant to expectancies for attaining a performance-approach goal.

Participants were led to believe that the experimental activity, a novel multiplication task, would be either simple or complex to learn, the latter to heighten the perceived challenge of the task. Those in a control group were given no expectations. After then completing measures of goal attainment expectancies and goal pursuit for the session, they learned the new multiplication technique and did one problem set using it. Performance on the problem sets and catch situational interest in the task, measured at the session's end, were the outcome measures.

Hypotheses

Task complexity perceptions should affect expectancies for attaining a mastery-approach goal but not for attaining a performance-approach goal. In particular, *Hypothesis 1* is that the expectancy for attaining a mastery-approach goal will be higher when the task is described as complex instead of simple. As the control group helps to anchor this comparison, we made no a priori hypothesis for how easy mastery-approach goal attainment will appear in the control condition relative to the two experimental conditions.

In Studies 1 and 2, neither of which could be classified as using either simple or complex course material, mastery-approach goals appeared easier than performance-approach goals. We expected in Study 3 to conceptually replicate this within the control condition.

Also, assuming that one's belief that a task is simple to learn will elevate mastery goal attainment expectancies (see Hypothesis 1 above), we expected the mastery-approach goal to appear easier than the performance-approach goal within the simple task condition too. This difference should be removed, and perhaps even become reversed, within the complex task condition. Thus, *Hypothesis 2* is that, within the control condition and the simple task condition alike, participants will report higher goal attainment expectancies for the mastery-approach goal than for the performance-approach goal.

Hypothesis 3 is that, regardless of condition, after removing shared variance between the two goal attainment expectancies, high mastery-approach goal expectancies will predict high pursuit of that goal but not the performance-approach goal, while high performance-approach goal expectancies will predict high pursuit of that goal but not the mastery-approach goal. *Hypothesis 4* is that the mastery-approach goal will again predict high catch interest, and *Hypothesis 5* is that the performance-approach goal will predict high performance on the activity. In sum, these hypotheses together test the links from (a) task perception to (b) goal attainment expectancies to (c) goal pursuit to (d) situational interest and achievement.

Method

Participants and design. The sample comprised 127 (76 female and 51 male) university students who participated in return for extra credit toward their course grade in general psychology. Approximately 72% were Caucasian and 28% were African American. Participants were randomly assigned to the three task description conditions in this between-subjects experiment.

Experimental activity, measures, and procedure. The experimenter, a student who remained blind to the study's purpose and hypotheses, began the session by explaining that the study involves learning and using a new technique for doing multiplica-

tion. Before learning the technique, participants were given 4 min to solve as many double-digit multiplication problems as they could using the traditional technique taught in schools. Their performance on this problem set served as an indicator of their baseline multiplication ability. They then completed a measure of their confidence doing multiplication ("I have a lot of confidence at doing problems like these" and "I'm just not good at doing problems like these" [reversed]; $\alpha = .89$) on the same *Strongly Disagree* (1) to *Strongly Agree* (7) Likert-type scale used for all self-report measures in this study. Baseline math ability and math confidence each served as covariates in all analyses.

Participants were then taught, via a folder of materials and an accompanying audiotape, a new technique for multiplying two-digit numbers (for details, see Barron & Harackiewicz, 2001). This "Left-to-Right" technique differs from the conventional technique taught in American primary schools in that it is faster and can be done mentally, without paper and pencil, thus making it useful for calculations in everyday day life (e.g., restaurant tips). The tape explained this in order to heighten the utility value of the task to all participants, regardless of experimental condition. The tape then delivered the manipulation. Participants in the *Control* condition were not told anything about the likely difficulty of learning the new technique. However, those in the *Simple Task* condition were informed by the tape:

This new technique is pretty simple to learn. It involves just four basic steps that you can do easily in your head. You won't even need to use paper and pencil for it. So you'll probably find the new technique straight-forward and easy to use right away.

Participants in the *Complex Task* condition were instead informed:

This new technique can be pretty challenging to learn. It involves doing four complicated steps in your head while also memorizing several large numbers. You do all of this without the aid of paper and pencil. So you might find the new technique hard to use at first.

The taped instructions then explained that participants would be given a set of 40 problems and 4 min to solve as many as they could using the new technique and that the experimenter would provide feedback about their score at the end of the session.

Participants then completed a single-item manipulation check ("The Left-to-Right technique seems like it will be pretty simple to learn"), plus measures of their expectancies for attaining mastery-approach goals ("I am confident I can learn the new multiplication technique"; "It will be difficult to master the new multiplication technique" [reversed]; $\alpha = .67$) and performance-approach goals ("I am confident I can do better than other participants in the study"; "It will be difficult to solve more problems than other participants" [reversed]; $\alpha = .78$). They then reported the degree to which they would pursue these achievement goals during the session. The mastery-approach goal pursuit measure emphasized skill development ("I would really like to master the new Left-to-Right technique"; "My goal in the session is to develop my skill with the new Left-to-Right technique"; "It is important to me to get better at using the Left-to-Right technique during the session"; $\alpha = .89$). The performance-approach goal pursuit measure emphasized normative success ("I would really like to do better than other participants in the study"; "My goal is to be able to solve

more problems than other participants"; "It is important to me to do better with this technique than other participants"; $\alpha = .87$).

Next, participants finished the tutorial of the new technique and did the problem set. The number correctly solved, out of 40, served as their task performance measure. Participants then received positive goal-relevant feedback from the experimenter. This feedback, presented on paper, provided the number of problems solved correctly, as well as goal-relevant information: in the mastery-approach goal condition, the form indicated that these scores represented "good" mastery of the task; in the performance-approach goal condition, it instead indicated that these scores represented "good" performance compared to previous participants, whose putative average score was several problems fewer than the participant's score. They then reported their Catch Interest, a five-item measure used in previous research with this activity (Barron & Harackiewicz, 2001; The left-to-right technique is "interesting," "fun," "enjoyable," "a waste of time" [reversed], and "boring" [reversed]; $\alpha = .87$).¹¹ Thus, *all* participants, whether pursuing the mastery-approach goal or performance-approach goal, received evidence of goal attainment before reporting their catch interest. This is important because mastery-approach goals naturally allow one to appraise goal attainment (i.e., a subjective sense of understanding the technique), whereas performance-approach goals require external feedback to make this appraisal (Dweck & Elliott, 1983). Standardizing competence perceptions therefore ensures that any goal effects on catch interest are due to the goal itself instead of the availability of goal-relevant feedback. Finally, participants were probed for suspicion about the feedback and purpose of the study; none expressed any on an open-ended questionnaire or during an exit interview.

Results

Preliminary analyses. As evidence of the manipulation's effectiveness, participants in the *Complex task* condition believed the new technique would be harder to learn ($M = 3.59$, $SD = 1.43$) than did participants in the *Simple task* condition ($M = 4.95$, $SD = 1.29$) or the *Control* condition ($M = 4.69$, $SD = 1.25$), $F(2, 124) = 12.21$, $p < .001$, $\eta^2 = .16$.

Primary analyses. Table 3 provides the means for each measure within each of the task description conditions. The only significant difference between conditions was for mastery-approach goal attainment expectancy, $F(2, 124) = 9.98$, $p < .001$, $\eta^2 = .14$. Bonferroni post hoc tests showed that, in accord with Hypothesis 1, participants judged the mastery-approach goal harder to attain when told the task was complex instead of simple or not given a task description. Furthermore, as expected, this description had no effect on their judgments of how hard the performance-approach goal would be to attain. No other differences between conditions were found.

Matching Hypothesis 2, paired t tests showed that goal attainment expectancies were higher for mastery-approach goals than performance-approach goals within the *Control* condition, $t(44) = 2.73$, $p < .01$, $d = 0.39$, and within the *Simple* condition, $t(42) = 4.54$, $p < .001$, $d = 0.59$. This pattern was reversed in the

¹¹ It is customary for laboratory studies with novel activities to assess "catch" interest instead of "hold" interest, as the latter is likely to take longer than a brief experimental session to develop.

Table 3
Condition Means for All Study 3 Measures

Variable	No description		Simple task		Complex task	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Mastery-approach goal attainment expectancy	4.84 _a	1.09	5.00 _a	1.16	3.95 _b	1.18
Mastery-approach goal pursuit	5.32	1.24	5.42	1.15	5.12	1.51
Performance-approach goal attainment expectancy	4.43	0.99	4.33	1.10	4.23	1.15
Performance-approach goal pursuit	3.90	1.47	3.97	1.38	3.59	1.49
Task performance	9.36	4.07	7.95	4.71	9.72	5.09
Catch interest	5.84	0.97	5.54	1.18	5.69	1.13

Note. Within each row, means that do not share a subscript differ significantly ($p < .05$).

Complex condition, $t(39) = 2.13$, $p < .05$, $d = 0.24$. These findings show how task perceptions shape beliefs about the *relative* difficulty of the two goals.

Antecedents of goal pursuit. Two sets of regression analyses were conducted to test the remaining two hypotheses. The first examined the antecedents to goal pursuit, in particular if mastery-approach and performance-approach goal attainment expectancies predict pursuit of the corresponding goal (Hypothesis 3). Because the task description manipulation affected mastery-approach goal attainment expectancies, we also wished to test if the manipulation directly or indirectly (via the expectancy) affects mastery-approach goal pursuit. Accordingly, we created two orthogonal contrasts to capture the three experimental groups (Cohen, Cohen, Aiken, & West, 2002). The first contrast, Task Complexity, pitted the Simple task (-1) against the Complex task ($+1$), the key comparison for our purposes. The second contrast, Task Description, pitted the No Task Description control condition (-1) against the Simple and Complex conditions ($+.5$ each). It tests the effect of simply having a description of the task's difficulty. The *Goal Antecedents Model* for the regression analyses included these two contrast codes, the two goal attainment expectancy terms, plus, as control variables, participants' gender, general confidence with math, and baseline ability.¹² Table 4 provides the zero-order correlations among all measures.

Mastery-approach goal. The overall model for mastery-approach goal pursuit was significant, $F(7, 119) = 7.32$, $p < .001$, $R^2 = .30$. Matching Hypothesis 3, participants were more likely to pursue a mastery-approach goal for the session if they held relatively high expectations for attaining it, $F(1, 119) = 10.14$, $p < .01$, $\beta = .35$. This expectancy effect also provided significant indirect pathways for general math confidence ($B = 0.05$), 95% CI $[.01, .14]$, and the Task Complexity contrast ($B = -0.15$), 95% CI $[-.26, -.06]$, to each influence mastery-approach goal pursuit, the latter such that simple tasks promoted greater mastery pursuit than complex tasks. Performance-approach goal expectancies were unrelated to mastery-approach goal pursuit.

Performance-approach goal. The overall model for performance-approach goal pursuit was significant too, $F(7, 119) = 5.00$, $p < .001$, $R^2 = .23$. Further supporting Hypothesis

3, participants were less likely to pursue a performance-approach goal if they believed it would be hard to attain, $F(1, 119) = 14.67$, $p < .001$, $\beta = .41$. Mastery-approach goal expectancies, however, did not predict performance-approach goal pursuit. Nor did general math confidence—either directly or indirectly through performance-approach goal attainment expectancies.

Consequences of goal pursuit. The second set of regression analyses tested if mastery-approach and performance-approach goals predict high catch interest and achievement, respectively (Hypotheses 4 & 5). These two educational outcomes were regressed onto a *Goal Consequences Model* comprising terms from the Goal Antecedents Model terms as well as mastery-approach goals and performance-approach goals. All of the possible Condition \times Goal interaction terms were nonsignificant in preliminary analyses and were therefore omitted.

Catch interest. The model for catch interest was significant, $F(9, 117) = 8.82$, $p < .001$, $R^2 = .40$. Matching Hypothesis 4, participants who more strongly pursued a mastery-approach goal later reported greater interest in the task, $F(1, 117) = 38.10$, $p < .001$, $\beta = .55$. Independent of this goal effect, participants also reported higher interest if they initially held lofty expectancies for being able to master the new technique, $F(1, 117) = 9.61$, $p < .01$, $\beta = .34$. This direct effect of expectancies was complemented by its positive indirect effect via mastery-approach goal pursuit ($B = 0.19$), 95% CI $[.09, .33]$. Finally, these same expectancies also provided a significant indirect pathway for the task description manipulation, such that simple tasks promoted higher interest than complex tasks ($B = -0.12$), 95% CI $[-.23, -.04]$.

Task performance. The model for task performance was significant, $F(9, 117) = 7.21$, $p < .001$, $R^2 = .36$. Matching Hypothesis 5, participants who more strongly pursued a performance-approach goal solved more problems correctly, $F(1, 117) = 3.96$, $p < .05$, $\beta = .19$. This goal effect also provided a significant indirect pathway for performance-approach goal attainment expectancy to influence task performance ($B = 0.08$), 95% CI $[.01, .13]$. Importantly, these effects were found even when controlling for math confidence, math ability, and gender, the latter two of which were significant predictors too; participants who solved more problems with the standard multiplication technique also solved more with the new technique, $F(1, 117) = 24.77$, $p < .001$, $\beta = .44$, and men solved more than women, $F(1, 117) = 22.73$, $p < .001$, $\beta = -.40$.

Model fit. Figure 3 diagrams the results described above. A SEM tested the overall fit. The covariates were allowed to correlate and, when significant (see Table 3), to predict the two goal attainment expectancy terms. The disturbance terms for the two goal attainment expectancies and for the two goals were also allowed to correlate due to shared method variance. This model fit the data adequately ($\chi^2/df = 1.68$, CFI = .92, RMSEA = .07, SRMR = .06).

¹² Preliminary analyses also included the interactions between each goal attainment expectancy term and the Task Complexity and Task Description terms, in order to test if the expected relationships between expectancies and the corresponding goal would vary by condition. None of those interactions was significant. They were therefore omitted from the final analysis in order to conserve power.

Table 4
Sample Means and Zero-Order Correlations Among All Study 3 Measures

XXX	1	2	3	4	5	6	7	8	9
1. Mastery-approach goal attainment expectancy	—								
2. Mastery-approach goal pursuit	.48	—							
3. Performance-approach goal attainment expectancy	.61	.29	—						
4. Performance-approach goal pursuit	.25	.38	.42	—					
5. Task performance	.10	.03	.20	.26	—				
6. Catch interest	.36	.55	.16	.13	.16	—			
7. General math confidence	.36	.18	.32	.28	.30	.12	—		
8. Baseline math ability	.12	.03	.10	.13	.37	-.04	.42	—	
9. Gender	-.18	.00	-.26	-.22	-.35	-.02	-.24	.18	—
<i>M</i>	4.62	5.29	4.33	3.83	8.99	5.69	5.75	17.65	
<i>SD</i>	1.22	1.29	1.07	1.44	4.64	1.40	1.40	6.32	
α	.67	.89	.78	.87	—	.87	.89	—	

Note. *N* = 127. For gender, 0 = male, and 1 = female. Correlations > .17 are significant (*p* < .05). α = internal reliability (Cronbach’s alpha) provided for all self-report measures using multiple items.

Summary

Study 3 supported each hypothesis. First, it shows that expectations for mastery-approach goal attainment are tethered to features of the activity (see also Church, Elliot, & Gable, 2001). Matching Hypothesis 1, participants held lower mastery-approach goal attainment expectancies when led to believe the new math technique would be complex and challenging to learn instead of simple. Furthermore, matching Hypothesis 2, participants in the control group held higher expectancies for attaining the mastery-approach goal than the performance-approach goal, thereby replicating Studies 1 and 2. The same pattern also emerged when the task was alleged to be simple. Thus, even when about to learn a novel technique for doing multiplication, a task that often arouses anxiety, participants still assumed it would be easier to learn and master the technique than to outperform peers with it. More

intriguing still is the fact that mastery-approach goal attainment was judged just as easy in the control condition as in the simple task condition. It appears that students consider mastery-approach goals relatively easy under baseline conditions.

These goal attainment expectancies matter, too. Matching Hypothesis 3, high expectancies for attaining a mastery-approach or performance-approach goal nudged students to pursue the corresponding goal but not the other goal. Thus, the task description had an enduring impact: indirect effects tests showed that merely describing the learning task as complex to learn had conspired to dampen mastery-approach goal pursuit. Furthermore, matching Hypotheses 4 and 5, mastery-approach and performance-approach goals predicted high catch interest and achievement, respectively, thereby revealing the distal impact of goal attainment expectancies. It is worth noting as well that these goal effects emerged even

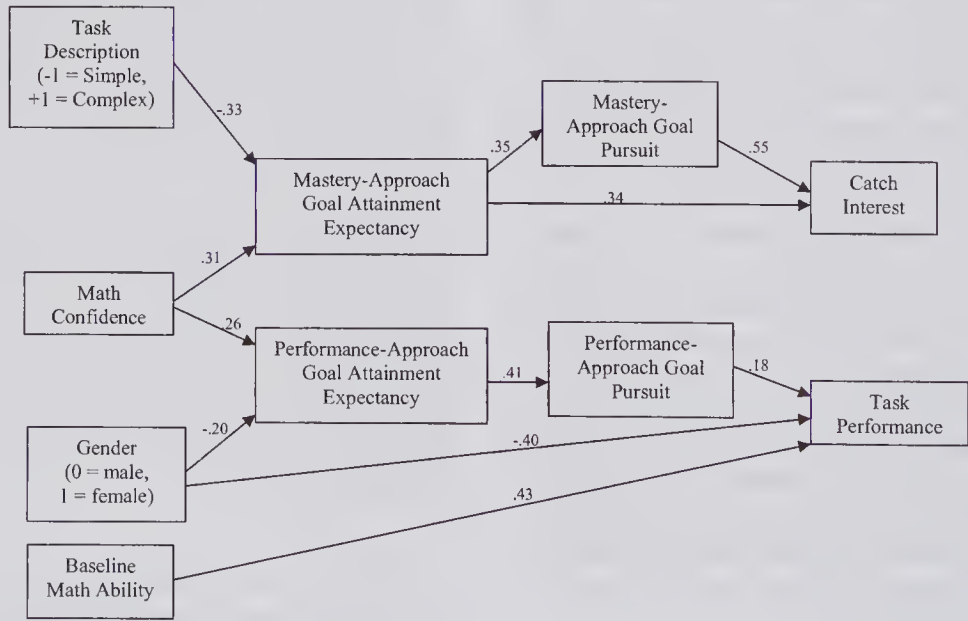


Figure 3. Full path model illustrating the goal antecedents and goal consequences pathways in Study 3. All paths are standardized regression coefficients (*p* < .05). For presentational clarity, nonsignificant paths and within-time correlations are omitted.

when statistically controlling for baseline math confidence, ability, and gender. The latter two did influence achievement—high ability aided achievement, and men solved more problems correctly than women, as in prior research with this math task (Barron & Harackiewicz, 2001)—but those individual differences cannot explain the performance-approach goal's link with achievement (Senko et al., 2011).

General Discussion

These three studies introduce the goal attainment expectancy construct to achievement goal theory. Despite minor differences in how they measured the key constructs, the studies revealed three consistent findings. One is that mastery-approach goals were generally judged easier to attain than performance-approach goals. This was evident across multiple general psychology classes (Studies 1 and 2) and also in the laboratory (Study 3); in the latter, it was shown not only when the math activity was alleged to be simple but even in the control condition in which no task description was offered. This adds generalizability to the initial findings of Senko and Harackiewicz (2005a), who found the same pattern with a word puzzle activity. It also fits Martin et al.'s (2008) finding that mastery-approach goals are more highly endorsed when the measure's items exclude challenge connotations.

Of course, goal attainment expectancies are not fixed. There surely are cases where mastery-approach goals appear quite hard to attain, perhaps even more so than performance-approach goals—for example, in classes where the material is complex and difficult to learn (e.g., science, technology, engineering, and math [STEM] courses; cf. Grant & Dweck, 2003). Study 3 demonstrated this. When the task was described as especially complex, goal attainment expectancies were lower for mastery-approach goals than performance-approach goals.

The second key finding is that, for each goal, the harder it appeared, the less likely people were to pursue it. This was evident in the present classroom and laboratory studies alike, and replicates a recent laboratory study by Jagacinski et al. (2010). Study 3 also offers initial support for our contention that goal attainment expectancies, compared to the broader competence expectancy belief, provide superior predictive value as a goal antecedent. Additional work is now needed to compare these constructs systematically across different learning contexts.

The third key finding is that, in each study, mastery-approach and performance-approach goals predicted situational interest and achievement, respectively. This, too, replicates past research (for a meta-analysis, see Hulleman et al., 2010) and therefore provides a reliable context in which to examine the goal attainment expectancy construct. Of course, the correlational nature of these findings makes causal interpretations challenging. In line with recent research (Harackiewicz et al., 2008; Van Yperen & Renkema, 2008), we believe that these relationships are probably recursive: That is, the two goals may promote achievement and interest, respectively, and these outcomes in turn may reinforce the continued pursuit of those goals.

Together, these three findings demonstrate that cues in the learning situation—such as task complexity and, we assume, peers' ability levels—can shape students' expectancies for attaining achievement goals, with meaningful consequences to their educational experience.

Limitations

The three studies have limitations, of course. Fortunately, in some cases, limitations to one study are absent from the others, and so the consistency between the three and with the broader goal literature allows reasonable confidence in the set of findings. For example, any ecological validity concerns with the laboratory method of Study 3 should be alleviated by the consistency in findings with the two classroom studies. Concerns about the directionality of the goal relationships with achievement in Studies 1 and 2 (cf. Brophy, 2005) should also be alleviated by the consistency with findings obtained in Study 3, which utilized a laboratory environment and controlled for general multiplication confidence and ability. Likewise, concerns about within-class dependence of observations in Studies 1 and 2 should be somewhat alleviated by the consistency between them and with the laboratory study. Finally, concerns about students' self-reported exam grades in Study 1 being inflated or invalid (see Kuncel, Credé, & Thomas, 2005) should be alleviated by the relatively low average grade ("C–") they reported, as well Study 2 showing the same effects with official course grades.

A consistent limitation in each study is that the goal attainment expectancy and goal pursuit measures were collected during the same session. We did this intentionally with the belief that goal expectancies are the proximal "in the moment" anchors to goal pursuit decisions. The alternative interpretation is that students who resolve to pursue a goal then decide that the goal must be easily attainable, as a way to rationalize their goal choice. If this were the case, however, then the task description manipulation in Study 3 should have directly affected mastery-approach goal pursuit. It did not. Instead, its effect was only indirect, by virtue of its impact on mastery goal attainment expectancy. This finding therefore supports the role of goal attainment expectancies as an antecedent, rather than outcome, of goal pursuit. Nevertheless, now that this link between goal attainment expectancies and goal pursuit is established, it may be useful for future research to further test how it unfolds over time.

The largest shortcoming of the current studies is that they focus solely on approach-based achievement goals. General competence expectancies often correlate positively with the two approach goals and negatively with the two avoidance goals (e.g., Bong, 2005; Elliot & Church, 1997). We expect a similar pattern for the two goal attainment expectancy constructs, except that each should correlate more with the approach and avoidant forms of the corresponding goal (e.g., mastery-approach expectancy predicting mastery-approach and mastery-avoidance goals) than with the other goal (e.g., the two performance goals). This remains for future work to test.

Along similar lines, our studies, like virtually all education-based research, relied on a broad and subjective definition of mastery-approach goals: to "learn," "master," or "develop skills." Yet these goals can also be defined concretely and objectively in terms of improving upon one's prior performance. With research now beginning to compare these two types of mastery-approach goals (Elliot, Murayama, & Pekrun, 2011), it may be useful to test if they differ in perceived ease or in the task features that affect their attainment expectancies.

Finally, the expectancy measures varied slightly between studies, and although the consistency in findings is reassuring, it would

be useful in future research to refine and systematically validate goal attainment expectancy measures. Existing measures of academic self-efficacy may provide a good starting point. Self-efficacy is one's judgment of being capable of succeeding at a task, and, as such, is often conceptualized as an antecedent to other motivational constructs, including goals (Bong & Clark, 1999). Two widely used measures of academic self-efficacy come from the Pattern of Adaptive Learning Survey (PALS; Midgley et al., 2000) and Motivated Strategies for Learning Questionnaire (MSLQ; Duncan & McKeachie, 2005). Inspection of their items reveals that, when used as a predictor of achievement goals, academic self-efficacy essentially serves as a measure of mastery-approach goal attainment expectancies: for example, "I'm certain I can master the skills being taught in this class" from the PALS, and "I'm confident I can understand the most complex material presented by the instructor in this course" from the MSLQ. This probably is why academic self-efficacy correlates higher with mastery-approach goals than performance-approach goals (for a review, see Payne et al., 2007). However, as Roeser (2004) observed, and the current findings demonstrate, the performance-approach goal's correlation would likely be stronger if the self-efficacy measure were to reference social comparisons instead of task mastery. Developing one as a companion to the traditional academic self-efficacy measure may prove advantageous for the integration of self-efficacy and achievement goal theories.

Theoretical Implications and Future Directions

Despite the above limitations, the present findings offer several contributions to achievement goal theory. One is that they refine theorizing about the interplay between competence perceptions and achievement goals. This interplay has historically been examined in two ways. Achievement goal theory originally treated competence perceptions as a moderator of any goal effects: Performance-approach goals should produce negative effects, and mastery-approach goals should produce positive effects, among students harboring self-doubts (Dweck, 1986). The other framework, later offered by Elliot and colleagues (Elliot & Church, 1997), treated *general* competence expectancy as an antecedent to mastery and performance goal pursuit. Their framework seems limited in predictive value, however, because general competence expectancies do not distinguish well between mastery-approach and performance-approach goals (e.g., Elliot & Church, 1997; Greene et al., 2004; Miller et al., 1993). This was also evident in Study 3, in which general math confidence, a suitable proxy for the broad competence expectancy construct, correlated modestly and *equally* with mastery-approach and performance-approach goals. Thus, the general competence expectancy construct distinguishes well between goal valences (approach vs. avoidance) but not goal contents (mastery vs. performance).

The current research offers a friendly amendment to Elliot's (1999) framework by substituting goal-specific attainment expectancies for general competence expectancies.¹³ Though the two goal attainment expectancies correlate with one another and may each be colored by general perceived competence, they are also, as the current studies show, easily molded by cues in the situation and have unique links with the corresponding approach goals. Prior work suggests they also negatively predict avoidance forms of each goal (Jagacinski et al., 2010). For these reasons, they provide

more precision and predictive value than do general competence expectancies.

Goal antecedents. Our findings also expand theorizing about the antecedents to achievement goals. They do so in two broad ways. The first is by introducing a new set of classroom-based antecedents. Theorizing on classroom-based antecedents has primarily focused on the classroom goal structure cultivated by teachers (Ames, 1992). Students are more apt to pursue performance-approach goals if their teacher emphasizes grades, underscores the role of ability in the learning process, or promotes social comparisons and ability attributions when evaluating students (e.g., Wolters, 2004). By contrast, students are more apt to pursue mastery-approach goals if their teacher emphasizes effort, makes the material personally meaningful, permits students some autonomy, uses individualized instead of normative criteria for evaluating students (Wolters, 2004), and fosters positive social relations with and among students (Patrick et al., 2011). Ultimately, these goal structure elements concern how a teacher's instructional technique, evaluation methods, and style of interacting with students can nudge them to pursue either performance-approach goals or mastery-approach goals. The current studies demonstrate that other cues in the learning context can also influence goal pursuit by shaping students' goal attainment expectancies. For instance, highly complicated material or confusing teachers might elevate the difficulty of a mastery-approach goal to the point of deterring its pursuit. Similarly, perceptions of peers' ability might affect whether someone pursues a performance-approach goal. From an application viewpoint, then, the current findings suggest that teachers who wish to cultivate a certain goal should consider not only traditional goal structure variables but also the students' beliefs about the attainability of the goal.

The second way in which our findings contribute to goal antecedents theorizing is by positioning the goal attainment expectancy construct as a conduit through which various antecedents affect students' goal choices. In particular, applying an expectancy-value type of framework, we propose that the power of goal antecedents is through shaping (a) how much students value the goal and (b) how much students expect to be able to attain the goal. Antecedents that fail to boost both are likely to have minimal impact on goal pursuit. For example, consider again how goal structures influence students' personal achievement goals. There is a curious pattern of findings: Mastery goal structures often correlate with personal mastery-approach goal pursuit more strongly than do performance goal structures with personal performance-approach goal pursuit (e.g., Anderman & Midgley, 1997; Bong, 2005; Friedel, Cortina, Turner, & Midgley, 2007; Gonida, Voulala, & Kiosseoglou, 2009; Kim, Schallert, & Kim, 2010). Furthermore, it appears that personal mastery-approach goals mediate the mastery goal structure effects on academic outcomes more reliably than personal performance-approach goals do for performance goal structure effects (Murayama & Elliot, 2009; Wolters, 2004). Presumably, a mastery goal structure and a performance goal structure are similarly potent in leading students to value the corresponding

¹³ We favored Elliot's (1999) framework here, but we believe that using two separate goal-specific competence perception constructs could also allow more precision with the earlier theoretical framework that treats competence beliefs as a moderator of goal effects.

personal goal but perhaps may have dissimilar impact on students' expectancies for attaining these goals. In particular, in contrast to a mastery goal structure, a performance goal structure likely conveys minimal teacher support and also implies a zero-sum element (i.e., only some students can succeed), both of which are likely to dampen some students' expectancies for goal attainment and, therefore, to deter their pursuit of the performance-approach goal.

Goal attainment expectancies may play a similar mediating role in how other antecedents affect goals. For example, it appears that students' interest has a recursive relationship with mastery goal pursuit, with baseline interest feeding mastery goal pursuit, which in turn strengthens interest (Harackiewicz et al., 2008; Hidi & Renninger, 2006). The conventional explanation is that interested students strive for this goal because they value task mastery. Yet, given that interest promotes academic self-efficacy (Silvia, 2003), it may be that interested students also strive for this goal in part because they have high expectancies for attaining task mastery. Expectancies may likewise help explain the impact of individual differences on students' goals. For example, need for achievement, test anxiety, competitiveness, and related traitlike qualities all appear to influence the achievement goals that students pursue (for reviews, see Baranik et al., 2010, and Moller & Elliot, 2006). Perhaps goal attainment expectancies are a proximal mechanism through which those traits impact goal pursuit. In sum, many studies have examined which classroom-based and student-based factors influence goal pursuit, typically with the tacit argument that these various factors lead students to *value* that goal. We propose that it may be fruitful to also consider how those factors shape students' *expectancies* for attaining that goal. The two elements may work together in meaningful ways.

Goal revision. The current studies focus on how initial goal attainment expectancies shape the *adoption* of achievement goals. This is merely a cross-section of an ongoing process. Students continually assess their goal progress, based on feedback from the learning environment, and decide whether to continue, table, alter, or abandon their goals. Their goal attainment expectancies are likely critical in this process (cf. Carver & Scheier, 1990; Schnelle, Brandstätter, & Knöpfel, 2010). Nicholls (1984) implied as much when theorizing that a drop in confidence might lead students to switch from a performance-approach goal to a performance-avoidance goal or disengage entirely, thus incurring various negative educational outcomes. Brophy (2005) later echoed this viewpoint when urging teachers to discourage performance goals in the classroom. This hypothesis has scarcely been tested but is supported in a pair of studies finding that poor task performance led students to simultaneously reduce their performance-approach strivings and increase their performance-avoidance (Senko & Harackiewicz, 2005b) or work-avoidance strivings (Kumar & Jagacinski, 2011), in both cases presumably because the negative feedback reduced performance-approach goal attainment expectancies. Curiously, theorists have spared mastery-approach goals of the same scrutiny. No one has argued that mastery-approach goals ought to be discouraged due to the possibility that low confidence in meeting the goal would lead to disengagement or a switch to mastery-avoidance goals. Yet the available data, though scant at this juncture, suggest that this could very well happen. Mastery-approach goals pursuit tends to be no more stable than performance-approach goal pursuit over the academic semester (see Senko et al., 2011), and each goal tends to be similarly

influenced by competence perceptions (Bong, 2005; Elliot & Church, 1997; Harackiewicz et al., 1997; Jagacinski et al., 2010; Senko & Harackiewicz, 2005b). The current studies demonstrate this: In each study, mastery-approach goal pursuit was predicted by mastery-approach goal attainment expectancies roughly as much as performance-approach goal pursuit was by performance-approach goal attainment expectancies. Furthermore, Study 2 showed that changes in mastery-approach and performance-approach goal attainment expectancies predict corresponding changes in mastery-approach and performance-approach goal pursuit, respectively. Thus, it appears that goal attainment expectancies may play a vital mediating role in goal revision, for mastery-approach and performance-approach goals alike. Future studies are needed to test this by including measures of the mastery-avoidance and performance-avoidance goals alongside their approach-based counterparts.

Conclusion

This article spotlights how students tether their pursuit of an achievement goal to their expectancies for attaining that goal. These expectancies, in turn, are likely to be shaped by and continually revised over time in response to goal-relevant cues in the broader learning context. We therefore recommend that theorists consider the potentially central role that these expectancies may play in the goal initiation and revision processes.

References

- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261–271. doi:10.1037/0022-0663.84.3.261
- Anderman, E. M., & Midgley, C. (1997). Changes in achievement goal orientations, perceived academic competence, and grades across the transition to middle-level schools. *Contemporary Educational Psychology*, 22, 269–298. doi:10.1006/ceps.1996.0926
- Arbuckle, J. L. (2007). *Amos* (Version 16.0) [Computer program]. Chicago, IL: SPSS.
- Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120, 338–375. doi:10.1037/0033-2909.120.3.338
- Baranik, L. E., Stanley, L. J., Bynum, B. H., & Lance, C. E. (2010). Examining the construct validity of mastery-avoidance achievement goals: A meta-analysis. *Human Performance*, 23, 265–282. doi:10.1080/08959285.2010.488463
- Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology*, 80, 706–722. doi:10.1037/0022-3514.80.5.706
- Bong, M. (2005). Within-grade changes in Korean girls' motivation and perceptions of the learning environment across domains and achievement levels. *Journal of Educational Psychology*, 97, 656–672. doi:10.1037/0022-0663.97.4.656
- Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist*, 34, 139–153. doi:10.1207/s15326985ep3403_1
- Brophy, J. (2005). >Goal theorists should move on from performance goals. *Educational Psychologist*, 40, 167–176. doi:10.1207/s15326985ep4003_3
- Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97, 19–35. doi:10.1037/0033-295X.97.1.19

- Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93*, 43–54. doi:10.1037/0022-0663.93.1.43
- Cohen, J., Cohen, P., Aiken, S. G., & West, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Covington, M. V., & Omelich, C. L. (1984). Task-oriented versus competitive learning structures: Motivational and performance consequences. *Journal of Educational Psychology, 76*, 1038–1050. doi:10.1037/0022-0663.76.6.1038
- Duncan, T. G., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist, 40*, 117–128. doi:10.1207/s15326985ep4002_6
- Dweck, C. S. (1986). Motivational processes affect learning. *American Psychologist, 41*, 1040–1048. doi:10.1037/0003-066X.41.10.1040
- Dweck, C. S., & Elliott, E. S. (1983). Achievement motivation. In P. H. Mussen (Series Ed.) & E. M. Hetherington (Vol. Ed.), *Handbook of child psychology: Vol. IV. Social and personality development* (pp. 643–691). New York, NY: Wiley.
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*, 169–189. doi:10.1207/s15326985ep3403_3
- Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York, NY: Guilford Press.
- Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72*, 218–232. doi:10.1037/0022-3514.72.1.218
- Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501–519. doi:10.1037/0022-3514.80.3.501
- Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology, 103*, 632–648. doi:10.1037/a0023952
- Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*, 5–12. doi:10.1037/0022-3514.54.1.5
- Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic self-concept: Reconsidering structural relationships. *Learning and Individual Differences, 19*, 499–505. doi:10.1016/j.lindif.2009.05.004
- Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: Sage.
- Friedel, J. M., Cortina, K. S., Turner, J. C., & Midgley, C. (2007). Achievement goals, efficacy beliefs and coping strategies in mathematics: The roles of perceived parent and teacher goal emphases. *Contemporary Educational Psychology, 32*, 434–458. doi:10.1016/j.cedpsych.2006.10.009
- Fryer, J. W., & Elliot, A. J. (2007). Stability and change in achievement goals. *Journal of Educational Psychology, 99*, 700–714. doi:10.1037/0022-0663.99.4.700
- Gonida, E. N., Voulala, K., & Kiosseoglou, G. (2009). Students' achievement goal orientations and their behavioral and emotional engagement: Co-examining the role of perceived school goal structures and parent goals during adolescence. *Learning and Individual Differences, 19*, 53–60. doi:10.1016/j.lindif.2008.04.002
- Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology, 85*, 541–553. doi:10.1037/0022-3514.85.3.541
- Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology, 29*, 462–482. doi:10.1016/j.cedpsych.2004.01.006
- Harackiewicz, J. M., Barron, K. E., Carter, S. M., Lehto, A. T., & Elliot, A. J. (1997). Predictors and consequences of achievement goals in the college classroom: Maintaining interest and making the grade. *Journal of Personality and Social Psychology, 73*, 1284–1295. doi:10.1037/0022-3514.73.6.1284
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94*, 638–645. doi:10.1037/0022-0663.94.3.638
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology, 92*, 316–330. doi:10.1037/0022-0663.92.2.316
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94*, 562–575. doi:10.1037/0022-0663.94.3.562
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, E. A., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100*, 105–122. doi:10.1037/0022-0663.100.1.105
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111–127. doi:10.1207/s15326985ep4102_4
- Hollenbeck, J. R., & Klein, H. J. (1987). Goal commitment and the goal-setting process: Problems, prospects, and proposals for future research. *Journal of Applied Psychology, 72*, 212–220. doi:10.1037/0021-9010.72.2.212
- Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin, 136*, 422–449. doi:10.1037/a0018947
- Jagacinski, C. M., Kumar, S., Boe, J. L., Lam, H., & Miller, S. A. (2010). Changes in achievement goals and competence perceptions across the college semester. *Motivation and Emotion, 34*, 191–204. doi:10.1007/s11031-010-9165-x
- Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review, 19*, 141–184. doi:10.1007/s10648-006-9012-5
- Kaplan, A., & Midgley, C. (1997). The effect of achievement goals: Does level of perceived academic competence make a difference? *Contemporary Educational Psychology, 22*, 415–435. doi:10.1006/ceps.1997.0943
- Kim, J. I., Schallert, D. L., & Kim, M. (2010). An integrative cultural view of achievement motivation: Parental and classroom predictors of children's goal orientations when learning mathematics in Korea. *Journal of Educational Psychology, 102*, 418–437. doi:10.1037/a0018676
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Kumar, S., & Jagacinski, C. M. (2011). Confronting task difficulty in ego involvement: Change in performance goals. *Journal of Educational Psychology, 103*, 664–682. doi:10.1037/a0023336
- Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75*, 63–82. doi:10.3102/00346543075001063
- Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70*, 647–671. doi:10.1177/0013164409355699

- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation. *American Psychologist*, 57, 705–717. doi:10.1037/0003-066X.57.9.705
- Marsh, H. W. (1990). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology*, 82, 623–636. doi:10.1037/0022-0663.82.4.623
- Martin, A. J., Marsh, H. W., Debus, R. L., & Malmberg, L. (2008). Performance and mastery orientation of high school and university/college students: A Rasch perspective. *Educational and Psychological Measurement*, 68, 464–487. doi:10.1177/0013164407308478
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, 82, 60–70. doi:10.1037/0022-0663.82.1.60
- Midgley, C., Maehr, M. L., Huda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., . . . Urda, T. (2000). *Manual for the Patterns of Adaptive Learning Scales*. Ann Arbor: University of Michigan.
- Miller, R. B., Behrens, J. T., & Greene, B. A. (1993). Goals and perceived ability: Impact on student valuing, self-regulation, and persistence. *Contemporary Educational Psychology*, 18, 2–14. doi:10.1006/ceps.1993.1002
- Moller, A. C., & Elliot, A. J. (2006). The 2×2 achievement goal framework: An overview of empirical research. In A. Mittel (Ed.), *Focus on educational psychology* (pp. 307–326). New York, NY: Nova Science.
- Muis, K. R., & Edwards, O. (2009). Examining the stability of achievement goal orientation. *Contemporary Educational Psychology*, 34, 265–277. doi:10.1016/j.cedpsych.2009.06.003
- Murayama, K., & Elliot, A. J. (2009). The joint influence of personal achievement goals and classroom goal structures on achievement-relevant outcomes. *Journal of Educational Psychology*, 101, 432–447. doi:10.1037/a0014221
- Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, 91, 328–346. doi:10.1037/0033-295X.91.3.328
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66, 543–578. doi:10.2307/1170653
- Patrick, H., Kaplan, A., & Ryan, A. M. (2011). Positive classroom motivational environments: Convergence between mastery goal structure and classroom social climate. *Journal of Educational Psychology*, 103, 367–382. doi:10.1037/a0023311
- Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology*, 92, 128–150. doi:10.1037/0021-9010.92.1.128
- Pintrich, P. R. (2000). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology*, 25, 92–104. doi:10.1006/ceps.1999.1017
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments & Computers*, 36, 717–731. doi:10.3758/BF03206553
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Renninger, K. A., & Hidi, S. (2011). Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist*, 46, 168–184.
- Roeser, R. W. (2004). Competing schools of thought in achievement goal theory? In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in research on motivation and achievement: Vol. 13. Motivating students, improving schools: The legacy of Carol Midgley* (pp. 265–300). Greenwich, CT: JAI Press.
- Ryan, A. M., & Pintrich, P. R. (1997). "Should I ask for help?" The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology*, 89, 329–341. doi:10.1037/0022-0663.89.2.329
- Schnelle, J., Brandstätter, V., & Knöpfel, A. (2010). The adoption of approach versus avoidance goals: The role of goal-relevant resources. *Motivation and Emotion*, 34, 215–229. doi:10.1007/s11031-010-9173-x
- Schunk, D. H. (1984). Self-efficacy perspective on achievement behavior. *Educational Psychologist*, 19, 48–58. doi:10.1080/00461528409529281
- Senko, C., & Harackiewicz, J. M. (2005a). Achievement goals, task performance, and interest: Why perceived goal difficulty matters. *Personality and Social Psychology Bulletin*, 31, 1739–1753. doi:10.1177/0146167205281128
- Senko, C., & Harackiewicz, J. M. (2005b). Regulation of achievement goals: The role of competence feedback. *Journal of Educational Psychology*, 97, 320–336. doi:10.1037/0022-0663.97.3.320
- Senko, C., & Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist*, 46, 26–47. doi:10.1080/00461520.2011.538646
- Silvia, P. J. (2003). Self-efficacy and interest: Experimental studies of optimal incompetence. *Journal of Vocational Behavior*, 62, 237–249. doi:10.1016/S0001-8791(02)00013-1
- Van Yperen, N. W., & Renkema, L. J. (2008). Performing great and the purpose of performing better than others: On the recursiveness of the achievement goal adoption process. *European Journal of Social Psychology*, 38, 260–271. doi:10.1002/ejsp.425
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81. doi:10.1006/ceps.1999.1015
- Williams, D. M. (2010). Outcome expectancy and self-efficacy: Theoretical implications of an unresolved contradiction. *Personality and Social Psychology Review*, 14, 417–425. doi:10.1177/1088868310368802
- Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236–250. doi:10.1037/0022-0663.96.2.236

Received July 29, 2011

Revision received November 2, 2012

Accepted November 7, 2012 ■

Constructing Motivation Through Choice, Interest, and Interestingness

Erika A. Patall
The University of Texas at Austin

Psychological research and theory have traditionally suggested that opportunities for choosing will lead to motivation and performance benefits. However, evidence on choice effects has not been ubiquitously positive, and recent investigations have revealed factors that diminish or reverse the effects of choosing. This investigation sought to extend this line of inquiry by examining whether interest factors may influence preferences for choosing and the effects of choice on motivation and performance. In Study 1, participants read a series of scenarios and reported a greater preference for choosing aspects of a task when the task was more, compared to less, personally interesting. Similarly, Study 2 revealed that choosing aspects of a trivia game enhanced post-task interest for the game only for individuals high in initial individual interest for trivia games in general. In contrast, Study 3 revealed that choosing enhanced post-task interest, perceived competence, value, and relative liking for a reading comprehension task when the reading passage was boring. When the passage was interesting, choosing resulted in less adaptive motivation outcomes. Going further, exploratory analyses revealed a 3-way interaction, suggesting that choosing enhanced willingness to engage in the task again only for those high in initial individual interest for reading and when the particular version of the task was boring. Interactions between choice and interest were not revealed for task performance in either Study 2 or Study 3. Rather, performance was higher among individuals who chose compared to individuals who did not. Implications of these findings are discussed.

Keywords: choice, autonomy, interest, motivation, self-determination

Supplemental materials: <http://dx.doi.org/10.1037/a0030307.supp>

Decades of psychological research suggest that all kinds of people (e.g., workers, the elderly, children), but students in particular, may feel more competent, more in control, more motivated, and perform better when they are able to express their preferences and make choices (e.g., Cordova & Lepper, 1996; Patall, Cooper, & Robinson, 2008; Patall, Cooper, & Wynn, 2010; Ryan & Deci, 2000). Further, the provision of choice is a common strategy used to motivate others in a variety of contexts (e.g., work, therapeutic, and educational). Teachers report that providing opportunities for choosing and decision making within the classroom or for school tasks is a popular method by which they attempt to enhance their students' motivation and learning (Flowerday & Schraw, 2000).

However, despite teachers' intuitive beliefs and the vast literature that exists, the controversy regarding the benefits versus detriments of choosing has yet to be put to rest. In fact, some studies find that choice may have no or even a negative effect on motivation and performance outcomes (Overskeid & Svartdal, 1996; Parker & Lepper, 1992; Reeve, Nix, & Hamm, 2003). A look at the literature on choice effects across a variety of outcomes suggests that there are likely both benefits and costs associated with making choices and that not all choices are equal for all

people or across all circumstances (Iyengar & Lepper, 1999, 2000; Moller, Deci, & Ryan, 2006; Patall et al., 2008; Reeve et al., 2003).

Prior research has suggested a number of factors (e.g., the type and nature of the choice, number of options given and choices made, or culture) that may influence the motivational benefits of choices (Iyengar & Lepper, 1999, 2000; Moller et al., 2006; Patall et al., 2008; Reeve et al., 2003). However, the motivational characteristics of the person and the task have yet to receive adequate attention as potentially important moderators of choice effects. That is, it seems reasonable to expect that the initial level of interest that an individual brings to a task is likely to influence how choice is experienced. By the same token, characteristics of the task such as its interestingness may also influence the effects of choosing on subsequent motivation and performance. It is a given circumstance that classrooms will often contain a heterogeneous population of students in terms of their motivational characteristics for various school tasks and that school tasks will necessarily vary in level of interestingness across students. Consequently, in order for choice provision to be most profitably used in educational settings, it seems imperative to assess the extent to which these factors influence the effectiveness of providing choice for enhancing students' motivation and learning outcomes.

The purpose of this investigation was to explore how interest influences preferences for making choices and the effect of providing choice on motivation and performance. The role of interest was explored in two ways. First, interest was investigated as a characteristic of the person approaching a task, as when an individual comes to a task with existing beliefs about how interesting

This article was published Online First October 8, 2012.

Correspondence concerning this article should be addressed to Erika A. Patall, Department of Educational Psychology, The University of Texas at Austin, 1 University Station D5800, Austin, TX 78712. E-mail: patall@austin.utexas.edu

and enjoyable he or she finds it based on a personal history of engagement and experience with the task. The role of interest as a characteristic of the environment was also examined, as when the task itself varies normatively across individuals in its ability to support a state of interest.

The Benefits and Detriments of Choice

According to self-determination theory (SDT), autonomy, competence, and relatedness are three fundamental needs that underlie people's intrinsic motivation, or the propensity to engage in a task for the inherent satisfaction it provides, and social contexts that satisfy these needs will enhance intrinsic motivation (Deci & Ryan, 1985; Ryan & Deci, 2000). Therefore, motivation is enhanced when contextual conditions allow people to feel that their actions are freely emanating from the self, afford people with the possibility of developing or demonstrating competence, and support a sense of belongingness with others in their environment. In contrast, when the environment is experienced as controlling, chaotic, and/or uncaring, psychological needs and intrinsic motivation are thwarted.

A great deal of research has supported the proposed positive effects of choosing, particularly in educational environments, demonstrating that the provision of choice leads to enhanced interest, enjoyment, effort, and persistence on a task (e.g., Cordova & Lepper, 1996; Iyengar & Lepper, 1999; Patall et al., 2008, 2010), as well as enhanced perceived competence, task performance, subsequent learning, preference for challenge, and creativity (e.g., Amabile, 1983; Cordova & Lepper, 1996; Iyengar & Lepper, 1999; Patall et al., 2008, 2010). Even neurological evidence has highlighted the inherent motivational quality of choosing, showing that people's anticipation of having choices is related to increased activity in the corticostriatal regions of the brain associated with reward processing (Leotti & Delgado, 2011).

Despite a great deal of theory and research suggesting that choice is a powerful motivator of behavior, not all studies have found choice to be a ubiquitously beneficial and some suggest it may even have a negative effect on adaptive motivation and performance outcomes (e.g., Flowerday & Schraw, 2003; Flowerday, Schraw, & Stevens, 2004; Overskeid & Svartdal, 1996; Parker & Lepper, 1992; Reeve et al., 2003). This complex pattern of previous findings beg the question, under what conditions does choosing lead to motivational benefits or detriments?

The Role of Interest in Choice Effects

Clearly, the effects of providing choices are complex and mixed findings suggest that there may be different effects of choice depending on the type of choice, the circumstances under which choices are provided, or the people making decisions. Past research has suggested a number of factors that are important to understanding these complex effects. The autonomy-supportive nature of the choice opportunity, the regulatory or cognitive demands of choosing, the number of options or opportunities for choosing, and the cultural background of the participant have all been found to be important moderators of the effects of choice on motivation and performance outcomes, among other factors (e.g., Iyengar & Lepper, 1999, 2000; Katz & Assor, 2007; Moller et al., 2006; Patall et al., 2008; Reeve et al., 2003).

Nevertheless, to this point, little attention has been given to the role of motivational characteristics of the person or the task in understanding when the provision of choice may be more or less beneficial. Of particular importance, "interest as a motivational variable refers to the psychological state of engaging or the predisposition to reengage with particular classes of objects, events, or ideas over time" (Hidi & Renninger, 2006, p. 112). Interest is often broadly conceptualized to include affective components (i.e., positive emotionality such as enjoyment) and cognitive components (i.e., evaluations related to continued engagement or reengagement; e.g., Hidi & Renninger, 2006; Krapp, 2002). In general, two forms of interest, individual and situational, have been identified in psychological and educational research to distinguish the momentary psychological state of interest from an enduring predisposition (e.g., Hidi & Renninger, 2006; Schraw, Flowerday, & Lehman, 2001). More specifically, *individual interest* (also referred to as personal interest) is a relatively stable disposition to reengage with particular content over time (cf. Hidi & Renninger, 2006; Schiefele, 2001). Individual interest primarily resides within the individual and refers to a general tendency to experience a psychological state of interest in reference to a particular content domain or class of activities. In contrast, *situational interest* refers to interest that primarily emerges from and is supported by the environment (Hidi & Renninger, 2006; Krapp, 2002). Situational interest is a momentary psychological state triggered by the environment (i.e., by the interestingness of the current content or activity) that may or may not last over time or re-occur when similar stimuli are presented. Indeed, years of research on interest as both a fleeting psychological state and an enduring disposition have suggested that interest supports an array of positive cognitive and behavioral outcomes (e.g., attention, persistence, engagement, and learning, among other outcomes; e.g., Ainley, Hidi, & Berndorff, 2002; Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Köller, Baumert, & Schnable, 2001; Renninger, Ewen, & Lasher, 2002; Schiefele & Krapp, 1996). Further, both forms of interest may influence the relations between choice provision and subsequent motivation and learning outcomes. However, the extent to which interest, either as an existing disposition or a state-like reaction to characteristics of the current environment, facilitates or diminishes the motivational qualities of choosing remains unclear.

Some theory and research has suggested that providing choices may buffer against the negative outcomes that poorly motivated students display and thus, providing choices may be particularly beneficial for those individuals who lack personal interest for the task at hand (Flowerday & Schraw, 2000; Schraw et al., 2001). While direct evidence on the role of interest in explaining choice effects is lacking, some support for this notion can be garnered from a phenomenological study of teachers' beliefs about instructional choice (Flowerday & Schraw, 2000), which found that teachers perceived choice to be especially beneficial for students who had low interest and little motivation for the task at hand. A complementary pattern was found in a study of German middle school students looking at class-to-class variation in perceived autonomy support and interest, where Tsai, Kunter, Ludtke, Trautwein, and Ryan (2008) found a stronger relation between perceived autonomy-support and daily interest for students with lower initial (individual) interest for the course subject.

Alternatively, some research has suggested a sensitization model in which optimally motivated students may benefit more than poorly motivated students from having the opportunity to make choices. Mouratidis, Vansteenkiste, Lens, and Sideridis (2011) found that elementary-age Greek physical education students with higher, compared to lower, relative autonomous motivation (i.e., motivation based to a great extent on interest, enjoyment, and value for course tasks) benefited significantly more from a need-supportive class in which teachers provided opportunities for making choices and working with other students.

Further and as previously alluded to, the role of interest in choice effects not only comes into play when considering individual differences in existing interest for a content area or activity (i.e., individual interest), but also when considering task interestingness. That is, the effects of choosing may be influenced by the extent to which situational interest is anticipated or experienced as a function of characteristics of the environment (i.e., the interestingness of the stimuli with which the individual is interacting with). On one hand, choice may be particularly powerful when a task is perceived as boring, as choosing provides an opportunity to build interest, enjoyment, and other forms of motivation during the task when little previously existed. Alternatively, choice may be more beneficial when the task at hand is perceived as interesting because people might be most receptive to factors that further influence their interest, enjoyment, or other aspects of motivation under this circumstance.

Accordingly, some motivation scholars have suggested that choice may be particularly motivating when it involves a task that is not interesting to begin with (e.g., Tafarodi, Milne, & Smith, 1999), although there is little evidence to examine this supposition. In line with this notion, it is worth noting that many demonstrations of improved motivation and performance due to choice have involved neutral or lackluster activities, such as solving anagram puzzles and paired-associate word learning (e.g., Iyengar & Lepper, 1999; Monty, Rosenberger, & Perlmutter, 1973; Perlmutter & Monty, 1973) or homework in a classroom context (Patall et al., 2010). Also providing some support for this possibility, Sansone, Weir, Harpster, and Morgan (1992) found that students who choose to make boring tasks more complex reported greater interest in those tasks.

Finally, it is worth pointing out that task interestingness and initial individual interest may not yield a parallel pattern of moderation and may interact in complex ways. That is, while the beneficial effects of choice seem likely to be most evident when an individual is interacting with a typically boring task (regardless of one's level of individual interest for the domain or class of activities), it does not necessarily follow that the benefits of choosing will be most pronounced among individuals with the parallel level of individual interest, that is, individuals with lower individual interest (e.g., see Mouratidis et al., 2011, as an example of research that would conflict with this hypothesis). In other words, there may be very different implications for the effects of choice depending on whether one considers on one hand, an individual's initial interest as an existing predisposition based on prior experiences with related tasks or on another hand, the experience or anticipation of situational (state) interest triggered by the interestingness of the task.

The Present Investigation

Providing students with choices appears to be a good strategy to support motivation and performance. However, there appears to be circumstances under which choosing may be more or less beneficial. Prior research has pointed to a number of factors that may influence the effects of choice. Nevertheless, limited attention has been paid to interest as a potential moderator of choice effects. To address this omission, a series of three experimental studies was conducted in which interest was either measured and/or manipulated in the context of the provision of choice.

First, to examine whether individuals would differ in their preference for having choice depending on their initial levels of individual interest, participants were asked to indicate the extent to which they would prefer to have task choices in response to a series of scenarios in which they imagined being asked to engage in a task under various conditions. Next, the effect of provision of choice on the motivation of individuals under various interest conditions was examined. In Study 2, participants' individual interest for the activity was measured prior to beginning it. In Study 3, the interestingness of the task was manipulated in addition to measuring participants' initial level of individual interest for the activity.

It was hypothesized that individuals would show a greater preference for making choices when their individual interest for the activity was higher. Likewise, it was expected that subsequent feelings of interest and enjoyment, competence, and other psychological benefits would be most enhanced by choosing among those who reported greater initial individual interest for the activity, but when the specific task was perceived as boring. Further, it was expected that individual interest, task interestingness, and the provision of choice might interact such that the benefits of choosing might be most dramatic when individuals came to a particular task perceived as boring with a high level of individual interest for the class of activity related to the one at hand.

Study 1

The investigation of whether interest influences the experience of choosing began by first examining whether people vary in their explicit preference to have the opportunity to make task-related choices under conditions of higher compared to lower interest. In Study 1, working adults and college students were asked to respond to two scenarios describing a situation in which they are asked to work on a task by either their boss or instructor. In one scenario, it was indicated that the actor had a high level of individual interest for the task at hand, while in the other it was indicated that the actor had a low level of individual interest. Participants' reports of their preference to make task-related choices served as the dependent measure.

Method

Participants. One hundred and fifty-two individuals (66% female) were recruited through Amazon.com's Mechanical Turk (MTurk), a website that allows researchers and businesses to post tasks and studies that the general public may peruse and participate in. Research on the use of MTurk have suggested that MTurk participants are more demographically diverse than are standard

Internet samples or typical American college samples and that the data obtained from MTurk are at least as reliable as those obtained via traditional methods (Buhrmester, Kwang, & Gosling, 2011). There were no exclusion criteria, and all participants who were over 18 years of age had an equal opportunity to participate in the study. Participants ranged between 18 and 65 years of age. The ethnic diversity of the sample approximated that of the U.S. population with the majority of participants identifying themselves as Caucasian ($n = 129$; 84%), eight participants identifying as Black (5%), 11 identified as Asian (7%), four identified as Hispanic (3%), and one identified as Native American (1%). Many participants were college students, but participants reported having a variety of occupations in a variety of fields. Participants earned a nominal contribution (\$1.50) to their Amazon.com account for participating in the study.

Procedure. All tasks associated with this study were completed online. Participants were informed that they would be presented with two scenarios describing a situation in which they are asked to work on a task. Participants were told to imagine that they were asked to engage in this task by either a boss or a course instructor. One scenario described a situation in which the participant was asked to engage in a personally interesting task, and the other scenario described a situation in which the participant was asked to engage in a personally boring task.¹ Participants were asked to think about each scenario and rate the extent to which they would prefer to have task-related choices given the circumstances described on a 5-point Likert scale ($1 = I$ would very much prefer not to make choices, $5 = I$ would very much prefer to make choices). The two scenarios were presented in a random order across participants. Once participants completed ratings of the extent to which they would prefer having choices, they were asked several background questions regarding their age, sex, ethnicity, and occupation.

Results and Discussion

In line with our hypotheses, results of a dependent t test indicated that participants had a greater preference for choosing in the situation in which they came to the task with high individual interest ($M = 4.26$, $SD = 1.07$) compared to the situation in which they came to the task with low initial individual interest ($M = 3.95$, $SD = 1.31$), paired samples $t(152) = .08$, $p = .33$; $t(151) = 2.92$, $p < .02$; $d = 0.32$.

The results of Study 1 suggested that overall, people reported neutral to positive attitudes toward having opportunities for making choices across both scenarios. However, results also provided initial support for the proposal that choosing would be more desirable under conditions in which the individual has greater individual interest for the task at hand. The results of this study thus inspired the question: Beyond explicit preferences for having choices, would an individual's initial interest for a task or the interestingness of a task influence the effects of choice on motivation and performance?

Study 2

Study 1 had suggested that people vary in their explicit preference for having choices depending on their individual interest, but how would this explicit attitude translate into motivation processes

and performance outcomes? To explore this question, college students completed a laboratory study in which they were asked to play a trivia and brain teaser game after having made choices or not about the topics of the puzzles. Initial individual interest for trivia games in general was assessed at the beginning of the study. Participants' performance on the trivia game and reports of their post-task interest and feelings of competence for the game served as dependent measures.

Method

Participants. Twenty-eight college students (19 females, 9 males) in several core psychology courses in a southern school were recruited to participate in the study. Students could participate in order to receive credit toward completing a research requirement for these courses. Participants ranged between 18 and 21 years of age. Participants were ethnically diverse: 18 participants were Caucasian (64.3%), one participant was African American (3.6%), eight were Asian (28.6%), and one was of mixed ethnicity (3.6%).

Procedure. Participants were run in individual experimental sessions. Aside from initial introductions and directions, all study activities were computerized. Participants were told that the purpose of the study was to investigate people's experience and impressions of a trivia and brain teaser game that the researchers had recently developed.

First, the trivia game was described to participants. Participants were told that they would be asked to complete 36 trivia and brain teaser questions of various types on all kinds of subjects. Examples were provided (e.g., Q: "What popular children's rhyme was an outgrowth of the bubonic plague?" A: "Ring around the Rosy"; Q: "What does x equal to solve the formula $(x + 1)(x - 1) = 0$?" A: " $+1$ or -1 "; Q: "What state can be spelled by rearranging the letters in the phrase: OLD FAIR?" A: "Florida"). Participants were told that they would have up to 45 s to answer fact-based questions and 2 min and 15 s for problem-solving and puzzle questions. If time ran out before the participant selected an answer, the computer automatically continued to the next screen. At this point, participants were asked to report on how interesting and enjoyable they generally find trivia games.

Next, participants were randomly assigned to complete the study under one of two choice conditions: *choice* or *no choice*. Participants in the choice condition were told that they would have three choices. Specifically, participants were told that although the computer would randomly select 36 questions from a bank of thousands of questions, participants had the opportunity to select three categories that they were guaranteed to receive questions on and from which the computer would over-select questions. For the first choice, participants were asked which category of trivia and brain teasers they would like to receive questions on among the following categories: (a) *Food and Drink*; (b) *History and Law*; (c) *Art, Literature, Entertainment, and Recreation*; (d) *People and Places*; (e) *Math and Science/Nature*; and (f) *Language, Riddles, and Puzzles*. For the second choice, they could choose a second category among the five remaining options. For the third choice, again, they chose a third category among the four remaining options.

¹ The interested reader can find transcripts of the scenarios in Appendix A of the online supplemental materials.

Participants in the no choice condition were made aware that there were several categories of question topics and that certain categories could be selected to be oversampled among the questions they would receive during the game. No choice participants were then told which topics had been assigned to them for the game. In reality, all participants were given the same set of trivia and brain teaser questions. Most questions fit into more than one category, so it was not obvious to participants that they were receiving just as many questions from non-selected categories as questions for selected categories. In order to further create the illusion that the choices had a real impact on the questions received, the order of the questions were arranged such that the last category chosen or assigned to participants was the first question to appear once they began playing the game.

The participant then worked on the 36 question trivia game for up to 40 min. After completing the game, the participant was asked to report on their perception of having choices, experience of interest, and feelings of competence in a post-task questionnaire. The number of questions the participant answered correctly also served as a dependent measure. Finally, participants were asked several background questions, including their sex, age, and ethnicity.

Yoking. A yoked design was used in which no choice participants were assigned the same categories of questions for the game that choice participants had previously selected. This yoking procedure allows participants in the choice condition choices while still ensuring that participants across conditions have the same task features. However, as previously mentioned, despite this protocol in which participants were led to believe they had chosen or been assigned particular question categories, there was no real difference in the actual game questions received. Participants were run intermittently through the choice and no-choice conditions to create 14 yoked dyads.

Materials. Interest-enjoyment and perceived competence subscales from the Intrinsic Motivation Inventory (IMI; Ryan, 1982) were adapted for use in this study. A version of the interest-enjoyment subscale was used to measure both initial individual interest for trivia and brain teaser games in general at the beginning of the study, as well as the post-task report of their experience of interest during *Brain Twister*. The perceived competence subscale was measured just once following engagement in *Brain Twister*. For the measure of initial individual interest, items were phrased in terms of trivia and brain teaser games in general (seven items; $\alpha = .93$; e.g., “I would describe trivia questions and brain teasers as very interesting,” or “I enjoy playing trivia games like *Brain Twister* very much”). For the post-task measures of interest

and perceived competence, items were framed in reference to the game that the participant had just completed (seven items for interest-enjoyment; $\alpha = .90$; e.g., “I enjoyed playing *Brain Twister* very much,” or “I would describe the game questions as very interesting”; six items for perceived competence; $\alpha = .92$; e.g., “I think I did pretty well on the game questions, compared to other college students”). Previous research has provided strong support for the validity (McAuley, Duncan, & Tammen, 1989) and reliability of this measure (e.g., Ryan, 1982). In addition, as a manipulation check, participants were asked about their perception of having received choices regarding the categories of questions for the game using four items explicitly designed for use in this study ($\alpha = .86$; e.g., “I believe I had some choice about the game questions I was given to complete”). Participants were asked to respond to all items using a Likert-type scale ranging from 1 (*not true at all*) to 7 (*very true*).

Results

Preliminary analyses. First, the distribution of scores on each variable was examined for statistical outliers. Grubbs’s (1950) test was applied, and no outliers were found.

Results suggested our choice manipulation was successful. Participants in the choice condition perceived having more choice regarding the question categories for the game ($M = 4.70$, $SD = 1.09$) compared to participants in the no choice condition ($M = 2.98$, $SD = 1.12$), $t(26) = 4.11$, $p < .001$, $d = 1.56$.

Motivation. To explore the proposal that the provision of choice in combination with initial individual interest would predict one’s subsequent experience of interest and perceived competence during the trivia game (“*Brain Twister*”), two hierarchical regression analyses (one for each outcome) were conducted. For each analysis, Step 1 included a dummy-coded variable to represent the choice manipulation (no choice = 0; choice = 1) and initial individual interest for trivia and brain teaser games. The interaction between these two variables was added at Step 2. All continuous predictor variables were centered using procedures detailed by Aiken and West (1991). Table 1 presents the correlations between the relevant variables, and Table 2 presents the results of these analyses.

The first step accounted for 53% of the variance in participants’ post-task reports of their experience of interest during *Brain Twister*, $F(2, 25) = 14.02$, $p < .001$. However, only initial individual interest in trivia and brain teaser games significantly predicted the experience of interest during *Brain Twister* ($\beta = .71$, $p < .001$), there was no main effect of choosing. The second step

Table 1
Means, Standard Deviations, and Bivariate Correlations Among Variables in Study 2

Measure	M (SD)	1	2	3	4	5
1. Provision of choice	0.50 (0.51)	—				
2. Initial individual interest	4.39 (1.12)	-.01	—			
3. Post-task interest	4.49 (1.00)	.14	.71***	—		
4. Perceived competence	3.34 (1.37)	.45**	.32*	.59***	—	
5. Task performance	11.85 (3.65)	.37*	.38**	.66***	.60***	—

Note. N = 28.
* $p < .10$. ** $p < .05$. *** $p < .001$.

Table 2
Regression Analyses for All Outcomes in Study 2

Predictor	Post-task interest			Perceived competence			Task performance		
	<i>B</i>	<i>SE</i>	β	<i>B</i>	<i>SE</i>	β	<i>B</i>	<i>SE</i>	β
Step 1									
Provision of choice	0.29	0.27	.15	1.21	0.45	.45**	2.65	1.22	.37**
Initial individual interest	0.64	0.12	.71***	0.40	0.20	.33*	1.24	0.55	.38**
Step 2									
Provision of choice	0.29	0.25	.15	1.21	0.46	.45**	2.64	1.22	.37**
Initial individual interest	0.32	0.19	.36*	0.30	0.34	.25	0.61	0.92	.19
Choice \times Individual Interest	0.50	0.24	.44**	0.15	0.43	.10	0.99	1.15	.24

Note. Provision of choice is dummy coded (0 = no choice condition, 1 = choice condition).

* $p < .10$. ** $p < .05$. *** $p < .001$.

contributed an additional 7% of the variance, an increase that was statistically significant, $F_{\Delta}(1, 24) = 4.32, p = .049$. This final model accounted for 60% of the variance in the experience of interest during the trivia game, $F(3, 24) = 12.02, p < .001$. The significant interaction between provision of choice and initial individual interest ($\beta = .44, p = .049$) was probed using simple regression equations of post-task interest on provision of choice at two levels of initial individual interest scores (see Figure 1). There was a significant positive effect of provision of choice on participants' post-task report of interest during the trivia game at two standard deviations above the mean of initial individual interest ($\beta = .71, p = .026$). In contrast, provision of choice did not significantly predict participants' reports of the experience of interest during the game at two standard deviations below the mean ($\beta = -.45, p = .16$).

To determine if a similar conclusion could be made for perceived competence on the game, this same analysis was conducted for that outcome. A different pattern of results emerged for perceived competence. The first step accounted for 30.7% of the variance in post-task perceived competence for the trivia game, $F(2, 25) = 5.53, p < .01$. There was a significant positive main

effect of the provision of choice on post-task perceived competence ($\beta = .45, p = .01$) and a marginally significant positive main effect of initial individual interest in trivia and brain teaser games ($\beta = .33, p = .06$). The addition of the interaction did not significantly contribute to the model, $F_{\Delta}(1, 24) = 0.13, p = .72$; total model $R^2 = .31$; $F(3, 24) = 3.60, p = .028$.

Task performance. To explore whether provision of choice in combination with initial individual interest would predict one's performance on *Brain Twister*, an identical hierarchical regression analysis for task performance as those previously described was conducted (see Table 2). A pattern of results similar to that for perceived competence emerged for task performance. The first step accounted for 28% of the variance in task performance, $F(2, 25) = 4.84, p = .017$. Similar to the analyses for perceived competence, both the provision of choice ($\beta = .37, p = .04$) and the initial individual interest for trivia and brain teaser games ($\beta = .38, p = .03$) significantly predicted task performance. The addition of the interaction did not significantly contribute to the model, $F_{\Delta}(1, 24) = 0.74, p = .40$; total model $R^2 = .30$; $F(3, 24) = 3.44, p = .03$.

Discussion

The results of Study 2 suggested that choosing provides motivational and performance benefits, especially for those individuals who had high initial interest going into the task. For individuals who entered the trivia game already with high individual interest for the activity, choosing led to enhanced feelings of interest for the current game compared to not choosing. In contrast, for individuals who entered the game having little initial interest in the activity, choosing had no effect on their subsequent experience of interest for the current game (and the non-significant effect was actually negative in direction). Surprisingly, this pattern of effect was not found for participants' feelings of competence and performance on the game. Rather, both the provision of choice and initial individual interest facilitated participants' perceptions of their competence on the game and their actual performance, but the effects of choice did not vary depending on participants' initial levels of individual interest for the activity.

These findings are compelling in that they seem to challenge the consensus, despite mixed results, that the provision of choice will unconditionally facilitate adaptive motivational and performance outcomes. This study seems to have helped to delineate one condition under which different effects of choosing may occur. In

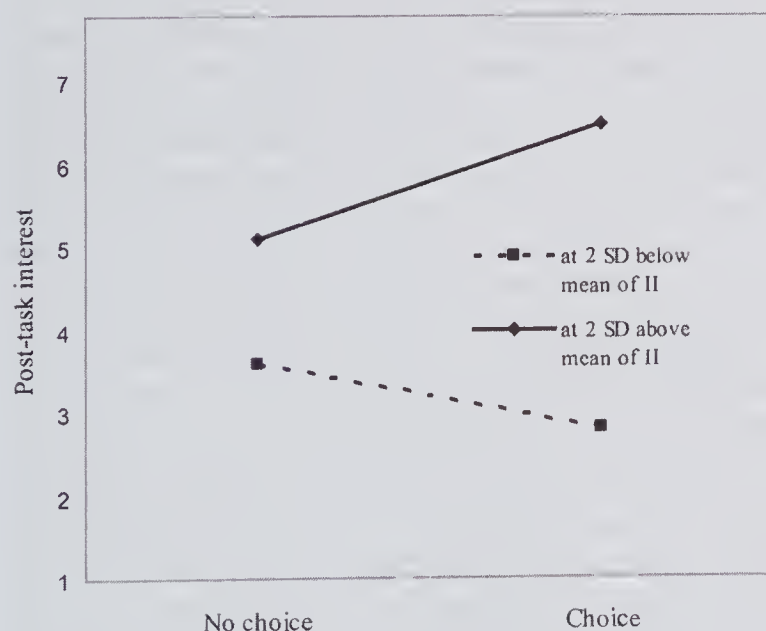


Figure 1. Regression of post-task interest on choice at 2 SDs above and below the mean of initial individual interest (II) in Study 2.

particular, this study suggests that for some motivation outcomes (the experience of interest and enjoyment), choice may only be empowering in the context of having some initial interest for the activity at hand. Providing and making choices may also be of some, but more limited value when an individual enters a task with low initial individual interest.

Study 3

Study 2 had suggested that the at least some of the motivational effects of choosing vary with people's initial individual interest. Given the findings of Study 2, one central question is whether differential effects of choice would be observed when interest was manipulated in the situation rather than measured as a characteristic of the individual. Further, given that in many cases, people will enter a task with both past experiences to inform their beliefs about how interesting they personally find an activity, as well as an understanding of how interesting the particular task at hand is, how these two factors might interact with choice to influence motivation and performance seemed to be an important question. To this end, Study 3 explored whether the effect of choice would vary when the interestingness of the task was manipulated and participants were informed of how interesting or boring most people had found the task in the past, as well as depending on people's initial individual interest for the activity at hand.

Method

Participants. One hundred and seventy-two college students (132 females, 39 males, 1 did not report) in several core educational psychology courses in a large southern school were recruited to participate in the study. Students could participate in order to receive credit toward completing a research requirement for these courses. Participants ranged between 18 and 26 years of age. Participants were ethnically diverse: 75 participants were Caucasian (44%), 18 participants were African American (10%), 23 were Asian (13%), 39 were Hispanic (23%), and 17 were another or of mixed ethnicity (10%).

Procedure. All tasks associated with this study were completed online using a commercial online survey software program, Qualtrics. Participants were informed that the purpose of the study was to investigate people's performance on a reading comprehension task under various conditions. Participants were told that they would receive a reading passage and several questions assessing their comprehension of the passage. After this description, participants were asked to report on how interesting and enjoyable they generally find reading.

Next, participants were randomly assigned to one of two task conditions: *boring task* or *interesting task*. At this point, participants were told that they would be receiving either an interesting or boring version of the reading comprehension task. Participants assigned to the boring task condition were told that they would be reading one of several articles on the scientific method from an academic text. Participants in this condition were told that most college students find the articles fairly boring. Participants assigned to the interesting task condition were told they would be reading one of several articles on employment among young professionals from a newspaper. Participants in this condition were told that most college students find these articles fairly interesting.

In essence, we varied the "interestingness" of the task by both informing participating students of how similar others have experienced the task and by selecting reading passages that we thought would be more or less personally relevant to interests of college students. Both boring and interesting articles were written at a ninth-grade reading level and were approximately the same length. At this point, a post-manipulation measure of interest for the upcoming reading comprehension task was taken to assess whether the task interestingness manipulation had influenced participants' perceptions of the task going into it (and prior to receiving any choices).

Next, participants were informed there would be several aspects of the task that could vary: the particular article and the difficulty of the questions. In the boring task condition, participants were told they could choose to read one of the following articles: "The Social Functions of Science" or "Teaching the Methods and Content of Science." In the interesting task condition, participants could choose to read one of the following articles: "Job Outlook Grim for Recent College Grads" or "Job Strategies Change in Challenging Economy." In addition, participants were told that the difficulty of the questions could vary such that all could be of medium difficulty or there could be a mix of easy, medium, and difficult questions. Participants were then randomly assigned to one of two choice conditions. In the *choice* condition, participants were then asked to make their choices for the reading comprehension task regarding the specific article they would read and the difficulty of the comprehension questions. Participants in the *no choice* condition were assigned these aspects of the task.

The participant then worked on the reading comprehension task for up to 15 min. The task consisted of a 420–450 word passage and seven reading comprehension questions. While participants had chosen or were assigned different articles and difficulty levels for the task, in reality there was no difference in the reading passages or questions within each task condition. Merely the title of the two articles differed so that choice participants could feel as if they had made a choice, without varying the task within task interestingness condition. Likewise, regardless of the difficulty level chosen or assigned to the participant, all participants received the same set of questions within each task condition.

After completing the reading comprehension task, the participant was asked to report on their perception of having choices, experience of interest and enjoyment during the task, perceived competence, the amount of effort they put into the task, their value for the task, their willingness to engage in the task again, and their relative liking of the task compared to similar ones. The number of questions the participant answered correctly also served as a dependent measure. Finally, participants were asked several background questions, including their sex, age, and ethnicity.

Yoking. As in Study 2, a yoked design was used in which participants were grouped into quads such that each member of a quad selected or received the same difficulty task option under one of the four experimental conditions and participants within each task interestingness category received the same article. In order to yoke participants across the two choice conditions (choice participants in the boring task condition and choice participants in the interesting task condition) and in turn, to participants in the no choice conditions, several participants in the choice conditions were run through the experiment first. A log of the choices each participant made was kept in order to determine when two choice

participants of varying task conditions naturally matched in the selection of difficulty assortment for the comprehension questions. Then, a matched participant in each of the no-choice conditions (one from both the easy and difficult task condition) was assigned options identical to those chosen by the participants in the choice conditions. Participants were run intermittently through the choice and no-choice conditions to create 43 yoked quads.

Materials. Measures identical to those used in Study 2 were also used in Study 3, with several additions and exceptions. Namely, in addition to the interest-enjoyment (initial: $\alpha = .93$; post-manipulation: $\alpha = .87$; post-task: $\alpha = .92$) and perceived competence ($\alpha = .92$) subscales of the IMI, subscales from the IMI measuring effort expenditure during the task (five items; $\alpha = .90$; e.g., "I put a lot of effort into this") and value for the task (seven items; $\alpha = .94$; e.g., "I believe the reading comprehension task could be of some value to me") were also measured in Study 3 following completion of the reading comprehension task. Consistent with Study 2, initial individual interest items were phrased to refer to reading in general and all post-manipulation and post-task items were worded to refer to the target reading comprehension task in the study. Like Study 2, participants were asked to respond to items using a Likert-type scale ranging from 1 (*not true at all*) to 7 (*very true*).

In addition, participants were asked in a single item to rate how willing they would be to work on the task again in the future using a Likert-type scale ranging from 1 (*not at all willing*) to 7 (*very willing*). Participants were also asked in a single item to indicate how much they enjoyed the reading comprehension task compared to similar tasks on a Likert-type scale ranging from 1 (*enjoyed it much less than other tasks*) to 7 (*enjoyed it much more than other tasks*). As with Study 2, participants were asked about their perception of having received choices regarding aspects of the reading comprehension task using the same four items explicitly designed for use in this investigation ($\alpha = .83$).

Results

Preliminary analyses. First, the distribution of scores on each variable was examined for statistical outliers. Grubbs's (1950) test was applied, and no outliers were identified. Results indicated that the choice and task interestingness manipulations were successful. Participants in the choice condition perceived having more choice regarding aspects of the reading comprehension task ($M = 4.59$, $SD = 1.15$) compared to participants in the no choice condition

($M = 3.24$, $SD = 1.01$), $t(170) = 8.10$, $p < .001$, $d = 1.25$. Likewise, participants in the boring task condition reported lower interest expectations for the upcoming reading comprehension task ($M = 2.85$, $SD = 1.20$) compared to participants in the interesting task condition ($M = 3.28$, $SD = 1.21$), $t(170) = 2.36$, $p = .02$, $d = -0.36$.

The effects of choice and task interestingness on motivation.

To explore the hypothesis that the interestingness of the task would moderate the effect of the provision of choice on one's subsequent motivation (i.e., post-task interest, perceived competence, effort, value, willingness to engage in the task again, and relative liking), a 2 (choice) \times 2 (task interestingness) between-subjects factorial multivariate analysis of variance (MANOVA) was conducted that included all six motivation outcomes (see Table 3 for means and standard deviations for each dependent variable by condition). Using Pillai's trace, the dependent variate was not significantly affected by the main effect of choice condition (Pillai's trace = .02), $F(6, 163) = 0.43$, $p = .86$. The multivariate main effect of task interestingness condition (Pillai's trace = .19), $F(6, 163) = 6.29$, $p < .001$, and the multivariate interaction between choice and task interestingness conditions were both significant (Pillai's trace = .13), $F(6, 163) = 4.09$, $p = .001$.

Univariate analyses of variance (ANOVAs) were conducted on each dependent measure separately to determine the source of the significant multivariate effects. Results suggested that individuals who received the interesting task reported feeling more interest during the task, $F(1, 168) = 25.48$, $p < .001$, $d = 0.74$; greater perceptions of competence, $F(1, 168) = 19.48$, $p < .001$, $d = 0.67$; greater value for the task, $F(1, 168) = 7.62$, $p = .006$, $d = 0.40$; greater willingness to engage in the task again, $F(1, 168) = 8.02$, $p = .005$, $d = 0.44$; and greater relative liking for the task compared to similar others, $F(1, 168) = 21.75$, $p < .001$, $d = 0.69$, compared to individuals who received the boring reading comprehension task. The main effect of task interestingness on effort was marginally significant, $F(1, 168) = 3.65$, $p = .06$, $d = 0.29$, and again, the pattern of findings indicated that individuals who received the interesting task put more effort into the task compared to individuals who had received the boring task. There was no main effect of choice condition on any motivation outcome.

More importantly, univariate analyses revealed a significant interaction between choice and task interestingness for post-task reports of interest, $F(1, 168) = 17.56$, $p < .001$; perceived competence, $F(1, 168) = 5.61$, $p = .02$; value, $F(1, 168) = 16.45$, $p < .001$.

Table 3
Means (and Standard Deviations) for Outcomes by Condition in Study 3

Dependent variable	Interesting task		Boring task	
	Choice ($N = 43$)	No choice ($N = 43$)	Choice ($N = 43$)	No choice ($N = 43$)
	M (SD)	M (SD)	M (SD)	M (SD)
Post-task interest	3.14 (1.10)	3.84 (1.14)	3.00 (1.02)	2.35 (0.97)
Perceived competence	4.51 (0.99)	4.83 (0.82)	4.20 (1.05)	3.80 (1.10)
Effort	4.27 (1.26)	4.60 (1.00)	4.25 (1.11)	3.94 (1.37)
Value	3.49 (1.24)	4.39 (1.17)	3.73 (1.19)	3.13 (1.25)
Willingness to engage	3.44 (1.52)	3.65 (1.77)	3.14 (1.44)	2.63 (1.38)
Relative liking	3.19 (1.28)	3.84 (1.29)	2.95 (1.19)	2.30 (1.21)
Task performance	3.37 (1.50)	3.12 (1.38)	3.58 (1.55)	2.93 (1.58)

.001; and relative liking, $F(1, 168) = 11.81, p < .001$. Tests of the simple effects, using the Bonferroni adjustment for multiple comparisons, revealed that among participants who had received the boring reading comprehension task, making choices significantly enhanced post-task reports of interest, $F(1, 168) = 8.30, p < .004, d = 0.65$; value, $F(1, 168) = 5.37, p = .02, d = 0.49$; and relative liking, $F(1, 168) = 5.91, p = .02, d = 0.54$, compared to individuals who did not make task choices. In contrast, among individuals who had the interesting task, making choices significantly diminished post-task reports of interest, $F(1, 168) = 9.28, p = .003, d = -0.63$; value, $F(1, 168) = 11.68, p = .001, d = -0.75$; and relative liking, $F(1, 168) = 5.91, p = .02, d = -0.51$, compared to individuals who did not make task choices. Although the pattern of effects was identical, the simple effect of choice condition on perceived competence was marginally significant among participants who had received the boring task, $F(1, 168) = 3.50, p = .06, d = 0.37$, and was not statistically significant among participants who had the interesting task, $F(1, 168) = 2.19, p = .14, d = -0.35$.

Looking at the simple effect of task interestingness for each choice condition, among individuals who had made task choices, there was no difference between those who had the interesting versus the boring task in terms of their post-task reports of interest, $F(1, 168) = 0.37, p = .54, d = 0.13$; perceived competence, $F(1, 168) = 2.09, p = .15, d = 0.30$; value for the task, $F(1, 168) = 0.84, p = .36, d = -0.20$; and relative liking of the task, $F(1, 168) = 0.75, p = .39, d = 0.19$. However, among individuals who had not made task choices, those who received the interesting task reported significantly greater interest, $F(1, 168) = 42.67, p < .001, d = 1.41$; perceived competence, $F(1, 168) = 23.00, p < .001, d = 1.06$; value for the game, $F(1, 168) = 23.24, p < .001, d = 0.64$; and relative liking of the task, $F(1, 168) = 32.81, p < .001, d = 1.23$.

The interaction effect was not statistically significant for effort, $F(1, 168) = 3.18, p = .08$, or willingness to engage in the task again, $F(1, 168) = 2.37, p = .13$. Nevertheless, the pattern of the means across conditions suggested a similar pattern of findings: Making choices enhanced effort ($d = 0.25$) and willingness to engage in the task again ($d = 0.36$) among individuals who had received a boring reading comprehension task, but diminished effort ($d = -0.29$) and willingness to engage in the task again ($d = -0.13$) among individuals who had received an interesting task.

The effects of choice and task interestingness on task performance. To explore whether provision of choice and task interestingness would predict one's performance on the reading comprehension questions, we conducted a 2 (choice) \times 2 (task interestingness) factorial analysis of variance (ANOVA) for task performance (see Table 3 for means and standard deviations by condition). The pattern of results that emerged for task performance was consistent with Study 2. Namely, there was a main effect of choice condition, $F(1, 168) = 3.91, p = .05, d = 0.31$, such that individuals who had made choices outperformed individuals who had not made choices. The main effect of task interestingness condition, $F(1, 168) = 0.003, p = .96, d = 0.01$, and the interaction between choice and feedback, $F(1, 168) = 0.74, p = .39$, were not significant. That said, the pattern of the means suggested that choice had a stronger positive effect on task

performance when participants worked on the boring ($d = 0.42$) compared to the interesting task ($d = 0.17$).

The interactive effects of choice, task interestingness, individual interest. To explore how the provision of choice in combination with task interestingness and initial individual interest would predict one's motivation and performance during and following the reading comprehension task, a series of hierarchical regression analyses (one for each outcome) was conducted that included provision of choice, task interestingness, initial individual interest, as well as all two- and three-way interactions.² The main interest of these analyses was the three-way interaction between choice, task interestingness, and initial individual interest for reading.

The three-way interaction between choice, task interestingness, and initial individual interest for reading was significant for only one of the seven variables examined, willingness to engage in the task again. The significant interaction between provision of choice, task interestingness, and initial individual interest ($\beta = -.28, p < .05$) was probed using simple regression equations of willingness to engage on provision of choice at two levels of initial individual interest scores and the two levels of task interestingness (see Figure 2). There was a significant positive effect of provision of choice on willingness to engage in the reading comprehension task again for participants who completed the boring task at two standard deviations above the mean of initial individual interest ($\beta = .51, p = .03$). Provision of choice did not significantly predict willingness to engage for participants who completed the boring task at two standard deviations below the mean ($\beta = -.27, p = .30$). Likewise, provision of choice did not significantly predict willingness to engage for participants who completed the interesting task at either two standard deviations above the mean ($\beta = -.31, p = .19$) or two standard deviations below the mean ($\beta = .13, p = .58$) of initial individual interest.

Discussion

The results of Study 3 suggested that choosing provides motivational and performance benefits particularly when the task is perceived as boring. For college students who were asked to engage in a boring reading task, choosing led to higher feelings of interest, value, and relative liking for the task compared to not choosing. In contrast, for college students who were asked to engage in a reading task perceived to be interesting for most college students, choosing had a negative effect on their motivation during and following the task. Looked at a different way, there was no difference in participants reports of their interest, perceived competence, value, or relative liking for boring and interesting versions of the reading task when choices were given. But, when choices were not given, participants reported significantly greater motivation when the reading task was interesting compared to when it was boring. In other words, choice seemed to create motivation where it did not exist such that boring and interesting versions of a task were experienced similarly. But, when choice was not present, the interesting task supported motivation far better than the boring task, not surprisingly.

² The interested reader can find tables presenting correlations between the relevant variables and results of these regression analyses for Study 3 in the online supplemental materials.

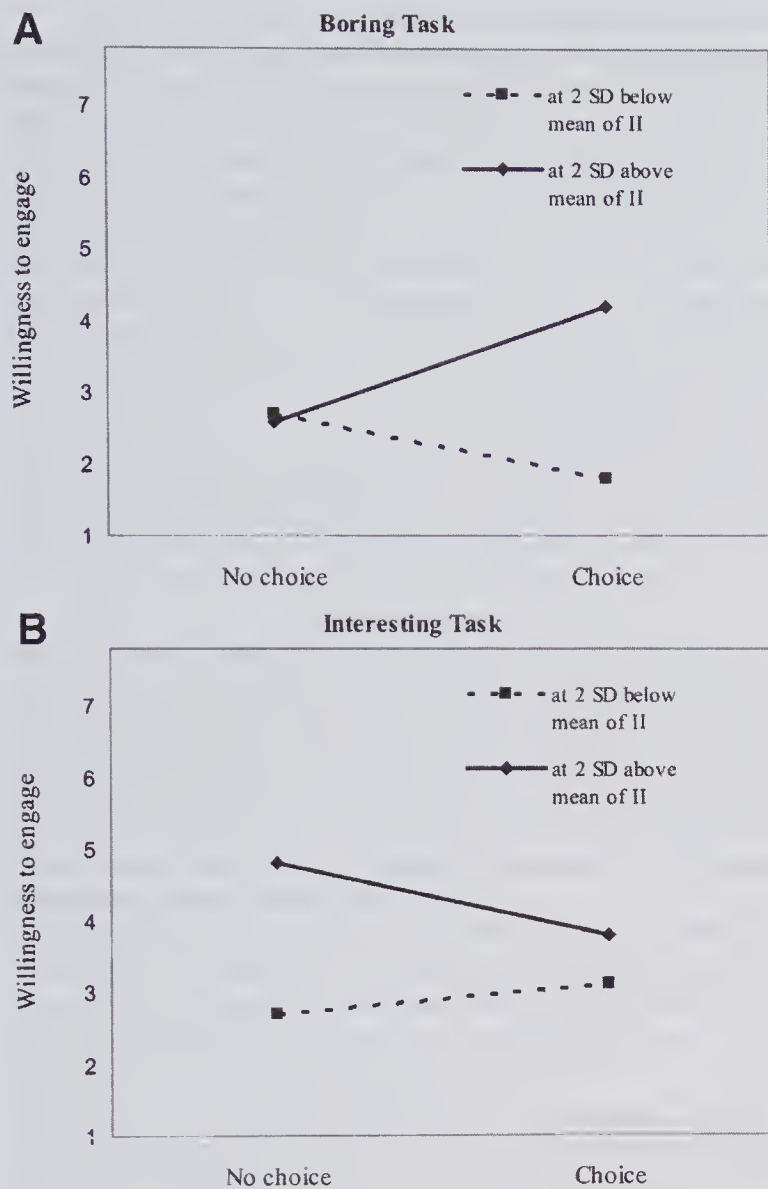


Figure 2. Regression of willingness to engage in the task again on choice at 2 SDs above and below the mean of initial individual interest (II) for boring and interesting tasks in Study 3.

Like Study 2, this pattern of effect was not found for participants' performance on the game. Rather, the provision of choice facilitated participants' actual performance on the reading comprehension questions, but the effect of choice did not vary depending on the interestingness of the reading task or participants' level of individual interest for the activity.

Exploratory analyses examining the three-way interaction between choice, task interestingness, and initial individual interest suggested that these three factors may interact in complex ways to affect motivation. In particular, the results of Study 3 suggest that choice is most facilitative of participants' willingness to engage in the task again in the future when an individual has high initial individual interest for the general activity, though the particular task is perceived to be boring. In fact, high initial individual interest in the context of a boring task was the only circumstance under which choice had a statistically significant effect on willingness to engage in the task again.

In sum, these findings help illustrate that choice will have different consequences for motivation depending on the conditions under which it is offered. This study suggests that choice may be

especially empowering in the context of a task that is perceived as uninteresting, and possibly, in combination with having high initial individual interest for the general activity at hand. Providing and making choices seems to be of more limited value when the task is perceived as typically interesting to most people.

General Discussion

While decades of psychological research have suggested that choice may generally lead to enhanced motivation and performance, especially among students in educational contexts, more recent investigations into the effects of choosing have challenged this assumption. Recent research on the effects of choosing has been fraught with mixed findings regarding the overall effect of choosing and has suggested that there are conditions under which and people for whom choosing may be more or less beneficial (e.g., Iyengar & Lepper, 1999, 2000; Moller et al., 2006; Patall et al., 2008; Reeve et al., 2003). The current findings help to provide a nuanced understanding of the conditions under which choice may be more and less beneficial. These are the first studies to demonstrate that offering an individual the opportunity to choose aspects of a task may be most beneficial when the individual feels some initial interest for the activity at hand or when the task is such that it can benefit from opportunities to build interest.

The present findings suggest that when individuals feel high compared to low individual interest for an impending task, they may have a greater preference for choosing and making choices further enhances their motivation for the task compared to not choosing. In fact, the enhanced benefits of choice in the context of high individual interest were found across three studies, despite the different methods of examining the questions. Thus overall, the results of this investigation seemed to support a sensitization model in which optimally motivated students, those with higher individual interest for the activity, seemed to benefit more than poorly motivated students from having the opportunity to make choices. These results suggest that for the individual with high individual interest, choosing may be experienced as desirable because it is an opportunity to maximize their potential to develop their skills, tailor the task to their particular preferences or goals, and perform successfully, while a lack choice may be seen as an unwarranted restriction of their ability to act autonomously, express their individuality, and maximize their skills. In contrast, for students who lack individual interest, choosing may be experienced as unnecessary, or even overwhelming. Rather than providing an opportunity to tailor the task to their personal preferences and goals, making task-related decisions may be an additional self-regulatory demand in the context of a task they already dislike.

That said, in some contradiction, results also support the notion that choice may lead to the greatest benefits for tasks that can stand to benefit from attempts to increase motivation outcomes (e.g., boring tasks). When considering the interestingness of the particular task rather than the individual's personal level of interest for the general activity, choosing seems to be especially beneficial in the context of a task that is perceived as uninteresting and potentially detrimental in the context of a task perceived to be interesting. This finding makes intuitive sense. Drawing on the notion of a ceiling effect, for a task that is already highly engaging, it may be more difficult to further increase one's motivation for that task. Further, an exploratory three-way interaction between choice, task

interestingness, and individual interest on participants' willingness to engage in the task again seemed to make the most sense of this apparent contradiction in the findings. Namely, Study 3 revealed that choice lead to significant enhancement in participants' willingness to engage in a boring task again when they started out with high initial individual interest for the general category of activity. But, choice had little benefit when the task was perceived to be typically interesting to most people or when people had low individual interest for the general activity.

In line with years of psychological research noting the motivational and performance benefits of intrinsic motivation, individual interest, and situational interest both in and outside of educational contexts and across various levels of schooling (e.g., Ainley et al., 2002; Harackiewicz et al., 2008; Krapp, 2002; Lepper, Corpus, & Iyengar, 2005; Ryan & Deci, 2000; Schiefele, 2001), initial individual interest and task interestingness were found in this investigation to have many benefits. In Study 2, individual interest for an activity predicted experiencing greater interest and enjoyment during task engagement, as well as heightened task performance. Likewise, in Study 3, interesting compared to uninteresting tasks led to the experience of greater subsequent interest and enjoyment during the target activity, as well as greater perceived competence, value for the task, willingness to engage in the task again, and liking of the task relative to similar ones. Given the highly engaging nature of the interesting compared to uninteresting task (as operationally defined in this investigation by the relevance of the information to the samples' personal goals and life concerns), it is little wonder that decision-making opportunities failed to further enhance motivational outcomes during an interesting task. Indeed, in the context of an interesting task, providing choices had negative motivational consequences. Perhaps in the context of an already interesting task, choice is experienced as a self-regulatory demand that has costs (decision-making effort) but few benefits.

While the two patterns of findings across Study 2 and 3 may appear in some contradiction to one other, two points might help to make sense of these findings. One resolution to this contradiction that has already been mentioned can be seen in the three-way interaction that was found in Study 3 between choice, interestingness, and individual interest for willingness to engage in the target task again. This interaction highlights that the three factors likely interact in complex ways and suggest that choosing may yield the greatest motivational benefits in particular contexts (e.g., when people both have some interest in the activity at the start and are given a particular version of that activity that is not naturally engaging). The fact that this three-way interaction was not found for other motivation outcomes is likely a result of the lack of power to detect such a complex interaction, given the relatively small to moderate effects (Cohen, 1988; Cohen, Cohen, West, & Aiken, 2003).

Second, this apparent contradiction highlights that all facets of interest do not operate equally. In these studies, individual interest was assessed by asking participants to think about how interesting and enjoyable they have generally found a category of activity in the past (i.e., trivia games in general or reading in general). Alternatively, instead of focusing on individual interest, anticipated situational interest (also a characteristic of the person) could have been assessed by asking participants to reflect on how interesting and enjoyable they expected the particular upcoming task to be after being given information about the task in order to make

such an assessment. This strategy may have led to different conclusions about how one's interest might moderate the effects of choice, conclusions that might be more in line to those which were drawn when task interestingness was manipulated. To further highlight this distinction, in contrast to how individual interest was measured, when manipulating the interestingness of the task, information about interestingness was provided in reference to the particular task (i.e., the particular articles to be read), rather than the category of activity in general (i.e., reading articles and answering comprehension questions in general). In other words, the specificity of the interest target may contribute to this apparent contradiction in the findings.

Somewhat surprising was the finding that the interactive effect of choice and interest factors was never revealed for task performance, and was only revealed for perceived competence in Study 3 in the context of the task interestingness manipulation. In both of the studies that tested the effects of choice on task performance (Studies 2 and 3), participants who made choices outperformed those who did not make choices about the tasks. Further, choosing had an impact on task performance even though there were no real differences in the task as a result of choosing. In Study 2, there was no difference in the trivia game questions that participants received even though they believed they had chosen categories of questions. Likewise, in Study 3, there was no difference in the reading passage or comprehension questions (within task interestingness condition) that participants received, even though participants thought they selected between two articles based on different titles and selected to receive questions of a particular difficulty assortment. Results suggest that there may be advantages of choosing that translate into differences in performance (and perhaps perceptions of competence) aside from its impact on motivation and emotion. Receiving one's preferences for aspects of a task, even when such preference matching is illusory or trivial, seems to yield cognitive processing benefits that result in enhanced performance.

Given the practical implications of choice-making effects in and outside the classroom, it seems imperative that future research replicate these findings and investigate whether the differential effects of choice observed in this set studies conducted primarily with college students and working adults applies equally to real life settings with various types of samples, especially pre-college students. Along these same lines, it remains unknown as to whether the conclusions of this set of studies might be generalizable to other tasks or other choice-making situations. It is possible that the relations between choice, task interestingness, and individual interest function differently when alternative tasks (tasks other than a trivia game or reading comprehension) are used or in the context of tasks that are not skill-dependent. A fruitful avenue of future research may be to investigate the effects of choosing and interest factors in other choice-making contexts, using other tasks and contrasting various targets of interest. Finally, in this research we have defined interest in line with how it is most commonly conceived of in educational psychology as including both an affective component (i.e., enjoyment) and cognitive components (i.e., perceptions of the activity having value and evaluations related to engagement or re-engagement; e.g., Hidi & Harackiewicz, 2000; Hidi & Renninger, 2006; Krapp, 2002). That is, our definition of individual interest and the situational experience of interest are hinged partially on the experience of enjoyment. It is important to note that this perspective is not unanimously agreed

on. While some emotion scholars agree that pleasantness needs to accompany interest (e.g., Ellsworth & Smith, 1988a, 1988b), others differentiate the positive emotions of interest and enjoyment, noting that they need not co-occur (e.g., Izard, 2007; Silvia, 2005; Turner & Silvia, 2007). In light of this controversy, future research could explore the roles of interest and enjoyment separately in explaining the effects of choice.

This research adds to the growing body of research demonstrating both the limits of choosing and the conditions under which choosing may be most valuable. Clearly, choice is to be valued for its ability to support some of the most important facilitators of learning. However, the provision of choice may need to be used judiciously, and in this case, used only after considering the level of individual interest of the person doing the choosing and the characteristics of the task the person is choosing about.

References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*, 545–561. doi:10.1037/0022-0663.94.3.545
- Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology, 45*, 357–376. doi:10.1037/0022-3514.45.2.357
- Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of cheap, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5. doi:10.1177/1745691610393980
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., West, S. P., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cordova, D., & Lepper, M. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology, 88*, 715–730. doi:10.1037/0022-0663.88.4.715
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press.
- Ellsworth, P. C., & Smith, C. A. (1988a). From appraisal to emotion: Differentiating among unpleasant feelings. *Motivation and Emotion, 12*, 271–302. doi:10.1007/BF00993115
- Ellsworth, P. C., & Smith, C. A. (1988b). Shades of joy: Patterns of appraisal differentiating positive emotions. *Cognition & Emotion, 2*, 301–331. doi:10.1080/02699938808412702
- Flowerday, T., & Schraw, G. (2000). Teacher beliefs about instructional choice: A phenomenological study. *Journal of Educational Psychology, 92*, 634–645. doi:10.1037/0022-0663.92.4.634
- Flowerday, T., & Schraw, G. (2003). Effect of choice on cognitive and affective engagement. *The Journal of Educational Research, 96*, 207–215. doi:10.1080/00220670309598810
- Flowerday, T., Schraw, G., & Stevens, J. (2004). The role of choice and interest in reader engagement. *The Journal of Experimental Education, 72*, 93–114. doi:10.3200/JEXE.72.2.93-114
- Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics, 21*, 27–58. doi:10.1214/aoms/1177729885
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100*, 105–122. doi:10.1037/0022-0663.100.1.105
- Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*, 151–179. doi:10.3102/00346543070002151
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111–127. doi:10.1207/s15326985ep4102_4
- Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology, 76*, 349–366. doi:10.1037/0022-3514.76.3.349
- Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*, 995–1006. doi:10.1037/0022-3514.79.6.995
- Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science, 2*, 260–280. doi:10.1111/j.1745-6916.2007.00044.x
- Katz, I., & Assor, A. (2007). When choice motivates and when it does not. *Educational Psychology Review, 19*, 429–442. doi:10.1007/s10648-006-9027-y
- Köller, O., Baumert, J., & Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education, 32*, 448–470. doi:10.2307/749801
- Krapp, A. (2002). An educational-psychological theory of interest and its relation to SDT. In E. L. Deci & R. M. Ryan (Eds.), *The handbook of self-determination research* (pp. 405–427). Rochester, NY: Rochester University.
- Leotti, L. A., & Delgado, M. A. (2011). The inherent reward of choice. *Psychological Science, 22*, 1310–1318. doi:10.1177/0956797611417005
- Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology, 97*, 184–196. doi:10.1037/0022-0663.97.2.184
- McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport, 60*, 48–58.
- Moller, A. C., Deci, E. L., & Ryan, R. M. (2006). Choice and ego-depletion: The moderating role of autonomy. *Personality and Social Psychology Bulletin, 32*, 1024–1036. doi:10.1177/0146167206288008
- Monty, R. A., Rosenberger, M. A., & Perlmutter, L. C. (1973). Amount and locus of choice as sources of motivation in paired-associate learning. *Journal of Experimental Psychology, 97*, 16–21. doi:10.1037/h0033784
- Mouratidis, A. A., Vansteenkiste, M., Lens, W., & Sideridis, G. (2011). Vitality and interest-enjoyment as a function of class-to-class variation in need-supportive teaching and pupils' autonomous motivation. *Journal of Educational Psychology, 103*, 353–366. doi:10.1037/a0022773
- Overskeid, G., & Svartdal, F. (1996). Effects of reward on subjective autonomy and interest when initial interest is low. *The Psychological Record, 46*, 319–331.
- Parker, L. E., & Lepper, M. R. (1992). Effects of fantasy contexts on children's learning and motivation: Making learning more fun. *Journal of Personality and Social Psychology, 62*, 625–633. doi:10.1037/0022-3514.62.4.625
- Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin, 134*, 270–300. doi:10.1037/0033-2909.134.2.270
- Patall, E. A., Cooper, H., & Wynn, S. R. (2010). The effectiveness and relative importance of providing choices in the classroom. *Journal of Educational Psychology, 102*, 896–915. doi:10.1037/a0019545

- Perlmutter, L. C., & Monty, R. A. (1973). Effect of choice of stimulus on paired-associate learning. *Journal of Experimental Psychology*, 99, 120–123. doi:10.1037/h0034749
- Reeve, J., Nix, G., & Hamm, D. (2003). Testing models of the experience of self-determination in intrinsic motivation and the conundrum of choice. *Journal of Educational Psychology*, 95, 375–392. doi:10.1037/0022-0663.95.2.375
- Renninger, K. A., Ewen, L., & Lasher, A. K. (2002). Individual interest as context in expository text and mathematical word problems. *Learning and Instruction*, 12, 467–490. doi:10.1016/S0959-4752(01)00012-3
- Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 43, 450–461. doi:10.1037/0022-3514.43.3.450
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78. doi:10.1037/0003-066X.55.1.68
- Sansone, C., Weir, C., Harpster, L., & Morgan, C. (1992). Once a boring task always a boring task? Interest as a self-regulatory mechanism. *Journal of Personality and Social Psychology*, 63, 379–390. doi:10.1037/0022-3514.63.3.379
- Schiefele, U. (2001). The role of interest in motivation and learning. In J. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement* (pp. 163–194). Mahwah, NJ: Erlbaum.
- Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text. *Learning and Individual Differences*, 8, 141–160. doi:10.1016/S1041-6080(96)90030-8
- Schraw, G., Flowerday, T., & Lehman, S. (2001). Increasing situational interest in the classroom. *Educational Psychology Review*, 13, 211–224. doi:10.1023/A:1016619705184
- Silvia, P. J. (2005). What is interesting? Exploring the appraisal structure of interest. *Emotion*, 5, 89–102. doi:10.1037/1528-3542.5.1.89
- Tafarodi, R. W., Milne, A. B., & Smith, A. J. (1999). The confidence of choice: Evidence for an augmentation effect on self-perceived performance. *Personality and Social Psychology Bulletin*, 25, 1405–1416. doi:10.1177/0146167299259006
- Tsai, Y., Kunter, M., Ludtke, O., Trautwein, U., & Ryan, R. M. (2008). What makes lessons interesting? The role of situational and individual factors in three school subjects. *Journal of Educational Psychology*, 100, 460–472. doi:10.1037/0022-0663.100.2.460
- Turner, S. A., & Silvia, P. J. (2006). Must interesting things be pleasant? A test of competing appraisal structures. *Emotion*, 6, 670–674. doi:10.1037/1528-3542.6.4.670

Received October 24, 2011

Revision received August 28, 2012

Accepted September 4, 2012 ■

Effectiveness of the KiVa Antibullying Program: Grades 1–3 and 7–9

Antti Kärnä
University of Turku

Marinus Voeten
Radboud University

Todd D. Little
University of Kansas

Erkki Alanen, Elisa Poskiparta,
and Christina Salmivalli
University of Turku

This study investigated the effectiveness of the KiVa Antibullying Program in two samples of students, one from Grades 1–3 (7–9 years old, $N = 6,927$) and the other from Grades 7–9 (13–15 years old, $N = 16,503$). The Grades 1–3 students were located in 74 schools and Grades 7–9 students in 73 schools that were randomly assigned to intervention and control conditions. Multilevel regression analyses revealed that after 9 months of implementation, the intervention had beneficial effects in Grades 1–3 on self-reported victimization and bullying (odds ratios ≈ 1.5), with some differential effects by gender. In Grades 7–9, statistically significant positive results were obtained on 5 of 7 criterion variables, but results often depended on gender and sometimes age. The effects were largest for boys' peer reports: bullying, assisting the bully, and reinforcing the bully (Cohen's d s 0.11–0.19). Overall, the findings from the present study and from a previous study for Grades 4–6 (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011) indicate that the KiVa program is effective in reducing bullying and victimization in Grades 1–6, but the results are more mixed in Grades 7–9.

Keywords: bullying, victimization, prevention, intervention, evaluation

Supplemental materials: <http://dx.doi.org/10.1037/a0030417.supp>

Can bullying among children and youth be reduced by school-based interventions? Despite some previous, somewhat pessimistic views (e.g., Ferguson, San Miguel, Kilburn, & Sanchez, 2007; Merrell, Gueldner, Ross, & Isava, 2008; J. D. Smith, Schneider, Smith, & Ananiadou, 2004), a recent meta-analysis by Farrington and Ttofi (2010) concluded that the intervention programs are effective in reducing bullying and victimization, with an average decrease of about 20% in the prevalence of these problems. There

was, however, considerable variability in results across studies, suggesting that the effectiveness of the programs may depend on the research methods, the nature of the intervention, and the target populations. The largest effects were obtained for intensive, long-duration programs with parent meetings and clear guidelines for tackling individual cases of bullying. It was also found that the effectiveness of programs increased steadily as the students got older (from 6 years to 14 years of age).

The results of Farrington and Ttofi (2010) concerning the influence of age are surprising and somewhat controversial. Specifically, several studies comparing the effects of one and the same program across age groups have shown that the programs actually work better for young rather than older students (Menesini, Codecasa, Benelli, & Cowie, 2003; Salmivalli, Kaukiainen, & Voeten, 2005) and better in primary than in secondary schools (Hanewinkel, 2004; Pitts & Smith, 1995; P. K. Smith & Sharp, 1994; Stevens, De Bourdeaudhuij, & Van Oost, 2000; see also Olweus, 2005, p. 4). All these studies were reviewed by P. K. Smith (2010, pp. 138–139), who concluded that antibullying programs often have less success in secondary than in primary schools. His main explanations were (a) developmental changes due to puberty and adolescence (e.g., in attitudes to victims) and (b) organizational changes resulting from larger and more complex structure of secondary schools. These organizational factors may make it more difficult to implement the intervention well.

As stated previously, the studies included in the meta-analysis by Farrington and Ttofi (2010) varied not only in the age of the target population but in other aspects as well. It is therefore important to continue studying the moderating effect of age while

This article was published Online First October 22, 2012.

Antti Kärnä, Department of Psychology, University of Turku, Turku, Finland; Marinus Voeten, Behavioral Science Institute, Radboud University, Nijmegen, the Netherlands; Todd D. Little, Department of Psychology and Schiefelbusch Institute for Life Span Studies, University of Kansas; Erkki Alanen, Elisa Poskiparta, and Christina Salmivalli, Department of Psychology, University of Turku.

The development of the KiVa program was financed by the Finnish Ministry of Education and Culture. The writing of this study was supported by the Academy of Finland Grant 134843 to the last author. We thank the whole KiVa project team for their contribution in the data collection process.

The University of Turku has the rights to disseminate the KiVa Antibullying Program. Because Christina Salmivalli and Elisa Poskiparta are among the authors of the KiVa material who stand to gain from a favorable report, they have distanced themselves from critical research activities such as primary data handling and analysis.

Correspondence concerning this article should be addressed to Antti Kärnä, Department of Psychology, University of Turku, Assistentinkatu 7, First floor, Turku 20500, Finland. E-mail: ankarna@utu.fi

using similar research methods and the same program across age groups. Furthermore, gender is another potentially important moderator of intervention effects, but the evidence on its role is scarce. The intervention effects on victimization have sometimes been larger for boys (Eslea & Smith, 1998; Olweus, 2004), whereas Olweus (2004) found larger reductions in girls' reports on bullying. In addition, age and gender are not only characteristics of individual students but also of classrooms. Students in a classroom form a social unit with a certain average age and a gender composition. It is quite possible that the social context in the form of these classroom-level characteristics strengthens or weakens the intervention effects on individual students. To our knowledge, however, in previous studies on antibullying programs, contextual effects have not been studied as moderators of intervention effects. Finally, in addition to positive results, there have been several interventions with statistically nonsignificant effects and even one with negative effects (Farrington & Ttofi, 2010). Despite the optimistic overall results of antibullying programs, it is therefore necessary to investigate the effectiveness of any new program when it is applied to its target population.

In the present article, we report the effects of the recently developed KiVa Antibullying Program on bullying, victimization, and other central outcome variables for Grades 1–3 and 7–9, thus extending the previous evaluation study for Grades 4–6 (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). We also compare the effects of KiVa for children in Grades 1–3, 4–6, and 7–9. The results of the present study provide new knowledge both on the effectiveness of the KiVa program and more generally on the effectiveness of antibullying programs on students, with age and gender taken into account both at the student and the classroom levels.

The KiVa Antibullying Program

The Finnish Ministry of Education and Culture funded the development and initial evaluation of a new antibullying program named KiVa (an acronym for *Kiusaamista Vastaan* [Against Bullying]). The program was developed at the University of Turku, in collaboration between the Department of Psychology and the Centre for Learning Research. The program was meant for elementary and lower secondary schools, and it was introduced in the intervention schools during two school years: first for Grades 4–6 (2007–2008) and 1 year later (2008–2009) for Grades 1–3 and 7–9, with another group of schools in the intervention condition.

Theoretical Background of the KiVa Program

KiVa is a theory-based intervention program with a background in a particular view of bullying and social behavior (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; Salmivalli, Kärnä, & Poskiparta, 2010a, 2010b). The program is based on (a) studies on the social standing of aggressive children in general (e.g., Cillessen & Mayeux, 2004; Rodkin, Farmer, Pearl, & Van Acker, 2000) and bullies in particular (Juvonen, Graham, & Schuster, 2003) and (b) research on participant roles in bullying (Salmivalli, Lagerspetz, Björkqvist, Österman, & Kaukiainen, 1996). At a more general level, social-cognitive theory (Bandura, 1989) is used as a framework for understanding the processes of social behavior.

Previous research suggests that bullying behavior is at least partly motivated by a pursuit of high status and a powerful position in the peer group (e.g., Juvonen & Galván, 2008; Salmivalli & Peets, 2008). Bullying can, in addition, be considered a group phenomenon, in which bystanders' behaviors have an effect on the maintenance of bullying and on the adjustment of the victims (Salmivalli, 2010; Salmivalli et al., 1996). Specifically, bystanders can contribute to the maintenance of bullying by assisting and reinforcing the bully, which provides bullies with the position of power; defending the victim, on the contrary, may make bullying an unsuccessful strategy for attaining and demonstrating high status (Salmivalli, Voeten, & Poskiparta, 2011). KiVa is predicated on the idea that a positive change in the bystanders' behaviors will reduce the rewards gained by bullies and consequently their motivation to bully in the first place. KiVa strongly emphasizes enhancing the empathy, self-efficacy, and antibullying attitudes of onlookers, who are neither bullies nor victims. This strategy is based on research relating these characteristics to defending and supporting victimized peers (Caravita, DiBlasio, & Salmivalli, 2009; Pöyhönen, Juvonen, & Salmivalli, 2010; Pöyhönen & Salmivalli, 2008; Salmivalli & Voeten, 2004). An important aim of KiVa is to make bystanders show that they are against bullying and to make them support the victim, instead of encouraging the bully. As another equally essential component, the KiVa program includes procedures for handling the acute bullying cases that come to the attention of the school personnel (for the program manuals, see Sainio et al., 2009, and Salmivalli, Poskiparta, Tikka, & Pöyhönen, 2009).

Universal interventions. The universal interventions of KiVa consist of three different age-appropriate versions that are, in the final version of KiVa, now widely implemented in Finnish schools, targeted at Grades 1, 4, and 7. During the evaluation study, these versions were introduced in Grades 1–3, 4–6, and 7–9, respectively. The universal interventions evaluated in Grades 1–3 were 10 double lessons for students (2×45 min each) given by classroom teachers during a school year. The lesson titles are "Let's Get to Know Each Other," "Emotions," "Our Class—Everyone Is Included!" "Difference Is Richness," "There Is No Bullying in KiVa School," "We Won't Join in Bullying!" "The Victim Needs Your Support," "I Will Not Be Bullied!" "Literature Lesson," and "KiVa Contract." The lesson goals are (a) to raise awareness of the role that the group plays in maintaining bullying, (b) to increase empathy toward victims, and (c) to promote children's strategies of supporting the victim and thus their self-efficacy to do so. The detailed lesson plans involve discussion, group work, role-play exercises, and short films about bullying. As the lessons proceed, class rules based on the central themes of the lessons are successively adopted one at a time. In the version now evaluated in Grades 7–9, four themes are described in the teachers' manual that can be introduced to students as series of lessons, whole theme days, or otherwise. The themes are "Group Interaction," "Me and the Others," "Forms of Bullying," and "The Consequences and Counterforces of Bullying." The recommended time to spend on the kick-off session, the four themes, and the concluding session compose 13–23 lessons altogether.

A unique feature of KiVa are the virtual learning environments involved. For primary school students (Grades 1–3 in the present study), there is an antibullying computer game that can be played during and between the student lessons. The game involves five

levels, and the topics and tasks in each level are closely connected to matters presented in the corresponding student lessons. By playing the game, students acquire new information and test their existing knowledge about bullying and learn new skills to act in constructive ways in bullying situations. Furthermore, they are encouraged to make use of these skills in real-life situations with their peers. For secondary school students, there is a different virtual learning environment called KiVa Street. It is an Internet forum where the students sign in and navigate to visit different places. For instance, they can go to a library and find information about bullying, or they can enter a movie theatre and watch short films about bullying. Similarly to the computer game, the KiVa Street aims to provide knowledge, skills, and motivation to change one's own behavior related to bullying. Both of these features thus form a component of the intervention in addition to those listed previously.

In all grade levels, KiVa provides prominent symbols such as bright vests for the recess supervisors to enhance their visibility and signal that bullying is taken seriously in the school and posters to remind students and school personnel about the KiVa program. Schools get presentation graphics they can use to introduce the program for the whole personnel and for parents. Parents also receive a guide that includes information about bullying and advice about what parents can do to prevent and reduce the problem.

Indicated interventions. In each school implementing the KiVa program, a team of three teachers or other school personnel, along with the classroom teacher, addresses each case of bullying that comes to their attention (Sainio et al., 2009; Figure 1). First, the team examines whether the reported case actually is an instance of bullying or not (e.g., a quarrel). The school team deals with bullying cases only; other conflicts are delegated to the classroom teacher. Second, individual discussions are organized with the victim. The victim gets a chance to relate his or her experiences, and the school team members communicate that they are on the victim's side and intend to put an end to bullying. This enhances the feeling of security for the bullied student. Third, each bully is taken, without prior notice, individually to discuss the bullying case. In this way, they do not have a chance to prepare themselves or to agree on a common story about the incidents. During the program evaluation phase, for research purposes, the school teams were randomized to implement one of two discussion methods: (a) a confronting approach, where the bullies are openly told that their behavior must stop immediately, and (b) a nonconfronting approach (cf. Pikas, 1989; Robinson & Maines, 1997), where the adult shares his or her concern about the victim and invites the bully to provide suggestions on what could improve the situation. Fourth, the school team meets with the bullies as a group to further confirm the agreements made individually. Fifth, there is a follow-up meeting with the victim to ascertain that bullying has stopped. An improvement in the situation is a requirement for the sixth phase, in which again a meeting is held with the bullies as a group. Also the victims may be included, if they want to be present. The goal of the meeting is to make sure that the bullying has stopped permanently.

In addition to the discussions with the involved students, the classroom teacher meets with between two and four prosocial and high-status classmates and encourages them to support the victimized child. For instance, this support may be shown by inviting the

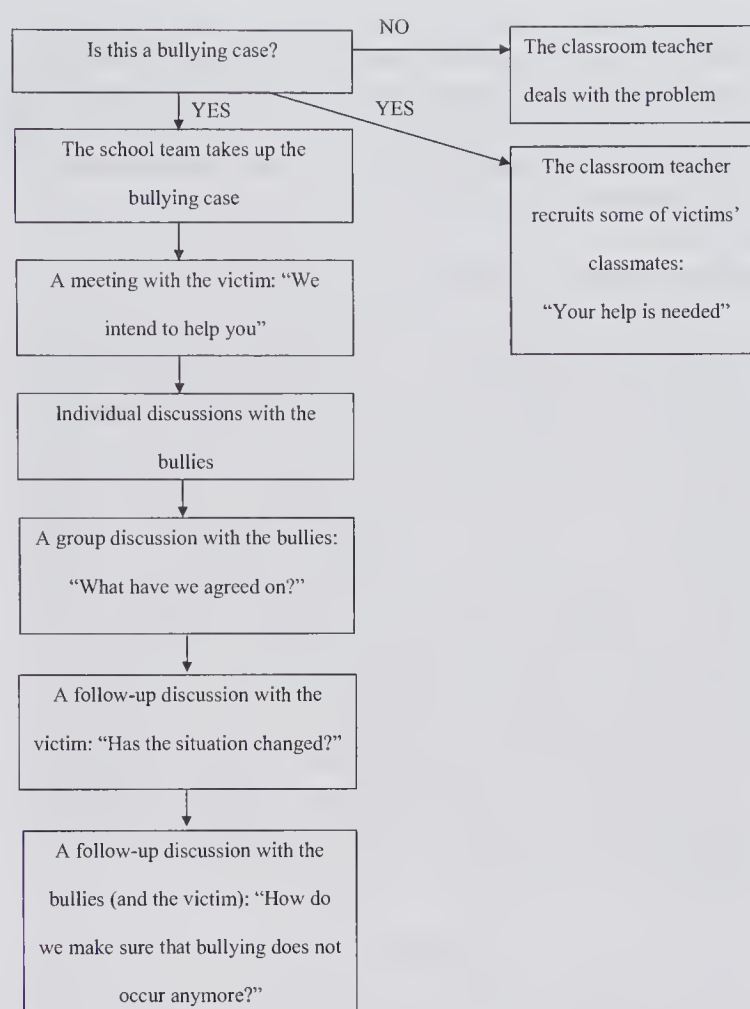


Figure 1. A flow chart for the individual and group discussions included in the indicated intervention.

victim in different activities, by treating the victim in an egalitarian and friendly way, or by trying to make others stop bullying.

Support in implementation. During the evaluation study, support was provided to teachers and schools to implement the program with fidelity. In addition to two full days of face-to-face training, networks of school teams were created, consisting of three school teams each. The network members met three times during the school year, with one person from the KiVa project guiding the network. The goal of the network meetings was to motivate the network members to implement the program and to help them overcome any possible obstacles in the process.

It is clear from previous description that KiVa is a whole-school antibullying program: Bullying is viewed as a systemic problem that has multiple causes at the student, classroom, and school levels (J. D. Smith et al., 2004). Like other whole-school programs (e.g., Olweus Bullying Prevention Program), the KiVa program includes components targeting individual students (e.g., discussion methods), classrooms (e.g., antibullying rules), and schools (e.g., a whole-school antibullying policy). There are, however, at least three features that taken together set KiVa apart from other antibullying programs (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). First, KiVa offers a comprehensive collection of concrete and professionally designed materials to be used in antibullying activities, not just abstract principles and guidelines. Second, the program makes use of modern technology

such as computer games and an Internet forum, which are engaging learning environments for the students. Third, while several antibullying programs emphasize the bystanders' role, the KiVa program provides concrete, research-based methods for enhancing the bystanders' empathy, self-efficacy, and efforts to defend the victimized peers.

Previous Studies on the Effects of KiVa

So far the effectiveness studies of the KiVa Antibullying Program have focused on Grades 4–6. The main findings can be summarized as follows: KiVa has been found effective in reducing bullying and victimization, and it has also reduced witnesses' negative behaviors (assisting and reinforcing the bully) and increased their self-efficacy to support and defend the victimized peers (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). The program effects have been found to generalize to multiple forms of victimization (Salmivalli, Kärnä, & Poskiparta, 2011). Furthermore, reductions in victimization in KiVa schools have been reported to predict decreases in depression, anxiety, and negative peer perceptions (Williford et al., 2012). Finally, the KiVa program has increased school liking, academic motivation, and self-reported academic achievement (Salmivalli, Garandeau, & Veenstra, 2012). All these effects were obtained during one school year (August to May) of implementation.

The Present Study

The effects of the KiVa program reported so far are promising. No study, however, has so far tested the effectiveness of KiVa on the main outcome variables among younger (Grades 1–3) or older (Grades 7–9) students in basic education. The present study offers new and important knowledge by examining the effectiveness of KiVa in new target populations. Furthermore, the study may shed light on the differential impact of antibullying programs on younger and older students, on boys versus girls, and on students in classrooms varying by average age and gender composition.

Similar to the researchers who conducted the previous study for Grades 4–6 (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011), we examined the program effects by comparing intervention-school students with control-school students at two time points: in the middle and at the end of the school year (i.e., 4 and 9 months after the beginning of program implementation; 7 and 12 months after the pretest measures). The baseline assessment took place at the end of the school year preceding the start of the intervention, because the intervention started right away at the beginning of the school year, and therefore, from a practical point of view, it was the latest possible date for a baseline assessment. The second measurement took place approximately halfway during the intervention year. This enabled us to investigate whether some intervention effects emerge already during the fall term.

We used self-reported bullying and victimization as the main outcomes; we expected that implementing the KiVa program would reduce these problems. For Grades 7–9, we also measured peer-reported bullying and victimization; reduction was expected for these outcomes in the intervention condition. For Grades 7–9, it was further hypothesized that the intervention would bring about beneficial changes in other outcomes. We expected (a) a decrease

in assisting and reinforcing the bullies and (b) an increase in defending the victims.

Method

Design and procedure. To recruit schools, we sent letters in the fall of 2006 to all 3,418 schools providing basic education in mainland Finland. These included both Finnish-language and Swedish-language schools, because the basic education in Finland is given in both official languages. The letter included information about the goals and content of the KiVa Antibullying Program and an enrollment form. The 275 volunteering schools were stratified by province and language, and 125 of them (excluding special-education-only schools) were randomly assigned to the intervention (47 schools) or the control condition (78 schools). Furthermore, 31 schools that had previously been randomized into the control condition for Grades 4–6 now participated in the intervention condition. This procedure resulted in a sample of 156 schools: 79 schools (40 control and 39 intervention) for Grades 1–3 and 78 schools (39 control and 39 intervention) for Grades 7–9. One control school participated both with Grades 1–3 and 7–9, but otherwise there was no overlap in the two samples. The Swedish-language schools were oversampled slightly (13% in the sample; 9% in the population). Because the participating schools were quite diverse and located throughout the country, they can be considered representative of those Finnish elementary and lower secondary schools that have an active interest in implementing the KiVa program.

The school year in Finland ranges from mid-August to the end of May. Data were collected in three waves: May 2008, December 2008–February 2009, and May 2009. Students filled out Internet-based questionnaires in the schools' computer labs during regular school hours. The process was administered by the teachers, who were supplied with detailed instructions about 2 weeks prior to data collection. The teachers were told to act in such ways that the confidentiality of the response was secured to a maximum extent, and also both younger and older students were assured that their answers would not be revealed to teachers or parents. In addition, teachers were offered support through phone or e-mail prior to and during data collection.

Teachers distributed individual passwords to the students, who used them to log into the questionnaire. At the beginning of the session, the term *bullying* was defined for the students in the way formulated in the Olweus' Bully/Victim Questionnaire (Olweus, 1996), which emphasizes the repetitive nature of bullying and the power imbalance between the bully and the victim. Compared with the original definition used in Grades 7–9, the definition for students in Grades 1–3 was shortened and simplified to facilitate understanding of the concept measured (see online supplementary appendix). Additionally, to remind the students of the meaning of bullying, a short version of the definition appeared on the upper part of the computer screen when the students responded to a bullying-related question. For Grades 1–3, the teacher read out loud the questions and the answering options in order to facilitate answering, whereas students in Grades 7–9 answered at their own pace. For the older students, the order of questions, items, and scales was extensively randomized to alleviate any systematic order effect. The sessions took on average 26 min in Grades 1–3 and 21 min in Grades 7–9 (5% trimmed means in May 2009).

The program implementation fidelity was measured in both of the samples. For Grades 1–3, the classroom teachers were asked to fill out a questionnaire immediately after each of the 10 KiVa lessons, whereas for Grades 7–9, the schools reported via a web-based questionnaire (in May 2009) about the activities during the intervention year. In this study, the implementation fidelity was represented as school-level averages of the number of given lessons and themes. Further details on this implementation process will be reported in upcoming publications.

Sample. The target sample for Grades 1–3 consisted of 79 schools (40 in intervention and 39 in control conditions). Two intervention and three control schools dropped out without providing any data at all, and therefore we ended up with a data set of 74 schools (38 intervention and 36 control; for details on drop-out schools in Grades 1–3 and 7–9, see online supplementary appendix). In these 74 schools, there were 7,739 students, of whom 7,231 (i.e., 93.4%) received active parental consent to participate in the study; 508 students were excluded from the analyses because of lack of parental consent. Another 304 students left the sample schools after the first wave of measurement, and they were excluded from the analyses, because they were not in the schools at the time of the intervention. This exclusion left us with a sample of 6,927 students in 397 classrooms in 74 schools to be included into the analyses. The number of students per classroom ranged from 1 to 30 ($M = 17.45$, $SD = 5.37$).

Parallel to Grades 1–3, the target sample for Grades 7–9 consisted of 78 schools (39 intervention and 39 control). Four control schools dropped out without providing any data, and one intervention school participated only in the first wave of data collection. After excluding these five schools from the analyses, we were left with 38 intervention schools and 35 control schools, in which there were 19,191 students. Of these students, 16,764 (i.e., 87.4%) gained active parental consent to participate. Altogether, 261 students left the sample after Wave 1 and were excluded. The final analysis sample consisted of 16,503 students in 1,000 classrooms in 73 schools. The number of students per classroom ranged from 1 to 26 ($M = 16.50$, $SD = 4.65$).

The youngest cohorts in our two subsamples were students in Grade 1 and Grade 7 during the intervention year. For these cohorts, we only had the posttest measurements, for two reasons. First, at the time of the pretest, these students were mostly not yet in the schools participating in the present study, and therefore it was impossible to collect pretest measurements from them. Second, we wanted to reduce the burden of data collection for schools participating in the study. Not all Grade 1 students would have had the required cognitive skills to respond to the questionnaire by Wave 2 (in December–February).

Due to the missing pretest measures for Grades 1 and 7, we fitted separate models for students at these grade levels. In the following, the reporting is focused on results for Grades 2–3 and 8–9; the posttest-only sample description and results for Grades 1 and 7 are summarized in the online supplementary material. In the final sample for Grades 2–3, there were 4,704 students in 273 classrooms in 74 schools, whereas in Grades 8–9, there were 11,070 students in 686 classrooms in 73 schools.

Classroom changes were not taken into account for either of the subsamples (Grades 2–3 or 8–9); in the fitted models, students were assigned at pretest to the classrooms they were in during the posttest measurements. This was done to keep the modeling task

tractable, and it was justified by the fact that the classroom student composition remained rather similar during the study. Comparing pre- and posttest measurements reveals that 80% of the classrooms remained at least 85% similar in Grades 2 and 3, whereas 90% of the classrooms remained at least 90% similar in Grades 8 and 9.

Missing data. There were several patterns of attrition in the data. For the independent variables, some values were missing, but they could be imputed on the basis of school records. The rates of unintentional missing data for dependent variables varied by variable and wave, but in general, the missingness was higher for self-reports (8.2%–18.4%) than for peer reports (3.2%–7.7%; for details, see Table A1 in the online supplementary material). For Grades 2 and 3, attrition was highest at posttest in the intervention schools, whereas for Grades 8 and 9, attrition was highest at posttest in the control schools.

There were students who responded at Wave 1, but whose answers were missing at Wave 3. Compared with Wave-3 responders, the largest Wave-1 differences were the following (Table A2 in the online supplementary material): The Wave-3 nonresponders had a higher level of some peer-reported behaviors—victimization (Cohen's $d = 0.11$), defending ($d = 0.08$), bullying ($d = 0.07$), and assisting the bully ($d = 0.06$)—and they had a higher level of self-reported bullying in Grades 2–3 ($d = 0.10$) as well as in Grades 8–9 ($d = 0.05$). For about half of the outcomes (five of nine), the drop-outs were at a disadvantage, but the differences were generally small. Additionally, in order to investigate differential attrition, we compared the intervention and control group differences at Wave 1 between the Wave-3 responders and nonresponders. The results indicated some potential for positive bias (i.e., inflation of intervention effects) in self-reported victimization in Grades 2 and 3, and in self-reported bullying and peer-reported defending in Grades 8 and 9. For other outcomes, the biasing effects of attrition were either small or negative. These mean comparisons do not reveal conclusively the mechanism of missingness (Enders, 2010), but they nevertheless suggest that missing data must be specifically taken into account in the models, and the results must be viewed with some caution.

Students could have one, two, or three measurements for a dependent variable. We treated these measurements as nested within students. The measurements were defined as Level 1 of the multilevel models. In this way, we could allow missing values at Level 1 and use all the available information, including the responses of students with partly missing data. The parameters of the models were estimated by full information maximum likelihood (FIML). Schafer and Graham (2002) considered this method as a state-of-the-art missing data technique (see also Jeličić, Phelps, & Lerner, 2009). This FIML approach works well (i.e., gives unbiased estimates) when the missing data can be assumed missing at random (MAR) and when the distributional assumptions for the residuals of the model are met (see, for instance, Enders, 2010, Chapter 4). Our analyses of the missing data showed differential attrition to some extent, and it is therefore unlikely that the missing data are missing completely at random. But MAR is a much less stringent assumption: MAR means that the probability of a missing value does not depend on the missing value itself but that this probability depends on the observed data used in the analysis model. This implies that all variables related to missingness need to be in the model. We can never know whether that is true, but it

is believed that FIML is rather robust to violations of the assumption (Collins, Schafer, & Kam, 2001).

Variables and instrumentation.

Self-reported bullying and self-reported victimization. The questionnaire started with demographic questions (e.g., gender and age) followed by questions about bullying and victimization. To measure bullying and victimization, we used the global items from the revised Olweus' Bully/Victim Questionnaire (Olweus, 1996): "How often have you been bullied at school in the last couple of months?" and "How often have you bullied others at school in the last couple of months?" Students answered with one of five frequency categories (0 = *not at all*, 1 = *only once or twice*, 2 = *two or three times a month*, 3 = *about once a week*, and 4 = *several times a week*). For the younger subsample, the five answering options were provided with different colors. That is, in addition to having the response options written in the web-based questionnaire, the teacher who was giving the instructions could also refer to the five colors when helping students pick the right alternative.

Students who reported they had been bullied two or three times a month, every week, or several times a week (Response Categories 2–4) during the past couple of months were categorized as victims, whereas those reporting they had bullied others at the same frequency were categorized as bullies. The cutoff point chosen agrees with the repetitive nature of bullying. With this cutoff point, victims and bullies differ markedly from noninvolved students in conceptually related variables (Solberg & Olweus, 2003). Furthermore, using this criterion (i.e., more than once or twice) facilitates comparisons between the present study and the previous studies, for instance those reviewed by Farrington and Ttofi (2010) in their meta-analysis. Dichotomization was also a practical way to deal with the extremely skewed distributions of the variables (Table 1). In addition, the scores on the self-reported bullying and victimization variables are only categorical or ordinal at most and clearly cannot satisfy the required distributional assumptions (e.g., normality).

To investigate the validity of the global bullying and victimization questions, we calculated school-level correlations between (a) the dichotomized global bullying and victimization items (used in the study), (b) averages for self-reported forms of bullying and victimization (e.g., name calling), and (c) peer reports of bullying and victimization (Kärnä, Voeten, Little, Poskiparta, Alanen, et al.,

2011). The results indicated that the associations between the global questions and the questions concerning the respective forms of bullying or victimization were substantial and fairly similar in all grade levels (Grades 1–9) of the KiVa data ($r_s = .65-.87$, $p < .001$). Also the correlations between the global bullying and victimization items were of similar magnitude across all grades ($r_s = .59-.65$, $p < .001$). In addition, for Grade Levels 4–9, school-level correlations between the global self-report items and peer-reported bullying and victimization were $.46-.75$ ($p < .001$). Taken together, these results provide clear evidence for the construct validity of the global bullying and victimization items. We chose to use only the global items because they provide unambiguous estimates for the prevalence of bullying and victimization. Averaging over ordinal responses concerning the various forms of bullying or victimization would have produced aggregate scores with no clear prevalence interpretation.

Participant roles in bullying situations and peer-reported victimization. When answering the Participant Role Questionnaire (Salmivalli & Voeten, 2004), students were instructed to think of situations in which someone was bullied. They were presented with items describing different ways to behave in such situations, and they were asked to nominate, from a list of classmates presented on the computer screen, an unlimited number of classmates who usually behave in the way described in each item. They were allowed also to choose "no one." The 12 items used in this study form four scales reflecting different participant roles: bullying ("Starts bullying," "Makes the others join in the bullying," and "Always finds new ways of harassing the victim"), assisting the bully ("Joins in the bullying, when someone else has started it," "Assists the bully," and "Helps the bully, maybe by catching the victim"), reinforcing the bully ("Comes around to watch the situation," "Laughs," and "Incites the bully by shouting or saying, 'Show him/her!'"), and defending the victim ("Comforts the victim or encourages him/her to tell the teacher about the bullying," "Tells the others to stop bullying," and "Tries to make the others stop bullying"). In order to measure peer-reported victimization, students nominated classmates treated in the following ways: "He/she is being pushed around and hit," "He/she is called names and mocked," and "Nasty rumors are spread about him/her" (Kärnä, Voeten, Poskiparta, & Salmivalli, 2010).

Table 1

Frequencies of Responses in the Five Categories of the Self-Reported Bullying and Victimization Variables at Wave 3

Variable	Grades 1–3				Grades 7–9			
	Victimization		Bullying		Victimization		Bullying	
	Frequency	%	Frequency	%	Frequency	%	Frequency	%
Occurrence								
Not at all	3,203	53.6	4,296	72.0	10,660	77.4	10,880	79.5
Only once or twice	1,745	29.2	1,333	22.3	2,031	14.7	1,987	14.5
2 or 3 times a month	446	7.5	197	3.3	402	2.9	344	2.5
About once a week	297	5.0	90	1.5	312	2.3	196	1.4
Several times a week	281	4.7	49	0.8	375	2.7	279	2.0
Participants								
Respondents <i>n</i>	5,972	100.0	5,965	100.0	13,780	100.0	13,686	100.0
Missing <i>n</i>	955		962		2,723		2,817	
Total <i>N</i>	6,927		6,927		16,503		16,503	

Peer nominations received were totaled and divided by the number of classmates responding, resulting in a score ranging from 0.00 to 1.00 for each student on each item. The proportion scores were averaged across the three items for each scale. The participant role scales have shown good internal consistencies (e.g., Salmivalli & Voeten, 2004), and in the present sample, Cronbach's alpha coefficients were .90 for the Bully scale, .86 for the Assistant scale, .83 for the Reinforcer scale, .88 for the Defender scale, and .73 for the Victim scale. The relatively low value of the latter is an exception, also compared with results from another sample (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). Finally, we examined the associations among these behaviors prior to intervention in Grades 7 and 8 ($n = 10,589$). There were strong positive correlations ($r_s = .76-.85$, $p < .001$) among the probullying behaviors (bullying, assisting, and reinforcing). The probullying behaviors had weak negative associations with defending (r_s ranging from $-.20$ to $-.26$, $p < .001$) but positive associations with victimization ($r_s = .12-.15$, $p < .001$). Defending correlated with victimization only weakly ($r = -.08$, $p < .001$).

Results

Descriptive Statistics for Outcome Variables

Before fitting the models, we examined the means and standard deviations, without statistical tests yet, for all dependent variables separately for the intervention and control groups at the three waves (Table 2). The comparisons revealed that, in general, there was an overall decrease in the mean levels and standard deviations of all dependent variables. These results suggest (a) that probullying and antibullying (i.e., defending) behaviors as well as victimization decreased over time and (b) that students became more similar over time. To some extent, the decreasing trends may also be a result of attrition (with more problematic students dropping out).

In pretest measures for Grades 2–3 and 8–9, the differences in averages between control and intervention groups were small (ranging from 0.00 to 0.01). For Grades 2 and 3, there was a clear positive change from Wave 1 to Wave 3 in the means of self-

Table 2
Descriptive Statistics for the Dependent Variables in Grades 2–3 and 8–9: Means and Standard Deviations

Variable	Control			Intervention		
	Wave 1	Wave 2	Wave 3	Wave 1	Wave 2	Wave 3
Grades 2–3						
Self-reported victimization						
<i>M</i>	0.23	0.16	0.17	0.22	0.13	0.13
<i>SD</i>	0.42	0.37	0.38	0.42	0.34	0.33
<i>N</i>	1,987	2,086	2,018	2,030	2,230	2,020
Self-reported bullying						
<i>M</i>	0.07	0.05	0.06	0.07	0.04	0.04
<i>SD</i>	0.25	0.23	0.23	0.26	0.20	0.20
<i>N</i>	1,966	2,083	2,018	2,027	2,224	2,019
Grades 8–9						
Self-reported victimization						
<i>M</i>	0.10	0.08	0.07	0.09	0.06	0.07
<i>SD</i>	0.30	0.27	0.26	0.29	0.24	0.25
<i>N</i>	4,333	4,360	3,847	5,694	5,535	5,252
Self-reported bullying						
<i>M</i>	0.08	0.06	0.07	0.07	0.06	0.05
<i>SD</i>	0.26	0.23	0.25	0.25	0.23	0.23
<i>N</i>	4,327	4,358	3,816	5,690	5,530	5,216
Peer-reported victimization						
<i>M</i>	0.07	0.06	0.05	0.06	0.06	0.05
<i>SD</i>	0.10	0.09	0.07	0.09	0.08	0.07
<i>N</i>	4,633	4,779	4,488	5,951	5,940	5,894
Peer-reported bullying						
<i>M</i>	0.05	0.05	0.04	0.05	0.05	0.04
<i>SD</i>	0.10	0.09	0.07	0.10	0.09	0.07
<i>N</i>	4,633	4,779	4,488	5,951	5,939	5,885
Peer-reported assisting						
<i>M</i>	0.07	0.06	0.05	0.07	0.07	0.05
<i>SD</i>	0.11	0.10	0.08	0.11	0.10	0.07
<i>N</i>	4,633	4,779	4,488	5,951	5,939	5,885
Peer-reported reinforcing						
<i>M</i>	0.11	0.11	0.08	0.12	0.10	0.07
<i>SD</i>	0.12	0.11	0.09	0.12	0.11	0.09
<i>N</i>	4,633	4,779	4,488	5,951	5,939	5,885
Peer-reported defending						
<i>M</i>	0.09	0.08	0.07	0.10	0.08	0.06
<i>SD</i>	0.10	0.10	0.09	0.10	0.10	0.08
<i>N</i>	4,633	4,779	4,488	5,951	5,939	5,885

reported victimization and bullying, whereas for Grades 8 and 9, the intervention/control differences in the mean-level changes were more mixed. They were in the positive direction for self-reported bullying and peer-reported reinforcing but in the negative direction (i.e., the change being more positive in the control group) for self- and peer-reported victimization and defending the victims.

Implementation Fidelity

We also examined whether the schools had actually used the KiVa program. To this end, we calculated school-level means and standard deviations for the implemented lessons and themes. In Grades 1–3, the teachers had given on average nine of the 10 prescribed lessons ($M = 9.1$, $SD = 1.1$; $N = 36$). The lower secondary schools were instructed to implement four themes plus introductory and concluding sessions, which amounts to six components. On average, the teachers had implemented five of the six prescribed components ($M = 5.1$, $SD = 0.9$; $N = 32$). It can be noted that the number of implemented lessons and themes corresponds well with the recommendations. This is, as one could expect, because the schools were participating on a voluntary basis.

Variances and Intraclass Correlations

We estimated the variance during pretest for Grades 2–3 and 8–9 (i.e., in the spring of Grades 1–2 and 7–8) for each dependent variable at three levels: students, classrooms, and schools (Table 3). There was nonzero variance for all variables at each level. We also calculated intraclass correlations (ICCs), which provide estimates of the proportion of variance due to differences among students, classrooms, and schools (for notation and formulas, see the note for Table 3). Classroom-level ICCs were higher for peer-reported (ICCs = .15–.25) than for self-reported (ICCs =

Table 3
Variance Estimates and Intraclass Correlations for Dependent Variables: Student (U), Classroom (V), and School (F) Levels

Variable	Variance			Intraclass correlations	
	$\hat{\sigma}_u^2$	$\hat{\sigma}_v^2$	$\hat{\sigma}_f^2$	ICC_1	ICC_2
Grades 2–3					
Self-reported victimization	1.304	0.156	0.074	.15	.05
Self-reported bullying	2.616	0.100	0.085	.07	.03
Grades 8–9					
Self-reported victimization	2.647	0.157	0.090	.09	.03
Self-reported bullying	3.812	0.241	0.071	.08	.02
Peer-reported victimization	0.734	0.186	0.025	.22	.03
Peer-reported bullying	0.667	0.099	0.015	.15	.02
Peer-reported assisting	0.745	0.127	0.020	.16	.02
Peer-reported reinforcing	0.803	0.205	0.044	.24	.04
Peer-reported defending	0.736	0.186	0.053	.25	.05

Note. All variances were statistically significant (one-tailed p -values at least $< .05$, but mostly $p < .001$), except for between-classroom variance for self-reported bullying in Grades 2 and 3 ($p = .059$). $\hat{\sigma}_u^2$ = variance between students; $\hat{\sigma}_v^2$ = variance between classrooms; $\hat{\sigma}_f^2$ = variance between schools; ICC = intraclass correlation. ICC_1 = proportion of total variance at the classroom level and the school level: $ICC_1 = (\hat{\sigma}_v^2 + \hat{\sigma}_f^2) / (\hat{\sigma}_u^2 + \hat{\sigma}_v^2 + \hat{\sigma}_f^2)$. ICC_2 = proportion of total variance at the school level: $ICC_2 = (\hat{\sigma}_f^2) / (\hat{\sigma}_u^2 + \hat{\sigma}_v^2 + \hat{\sigma}_f^2)$.

.07–.15) data. It should be kept in mind, however, that ICCs at the classroom level include both classroom- and school-level variance (classrooms are nested within schools). The highest proportions of variance associated with classroom or school factors were obtained for peer reports of defending ($ICC = .25$), reinforcing ($ICC = .24$), and victimization ($ICC = .22$). For all outcomes, the classroom-level variance was clearly higher than the school-level variance. Between-school variance was highest for peer-reported defending ($ICC = .05$), and, in Grades 2–3, for self-reported victimization ($ICC = .05$). Overall, these ICCs show that students sharing the same social environment were more alike than students from different classrooms or schools.

Multilevel Models

We used multilevel modeling with MLwiN Version 2.22 (Rasbash, Charlton, Browne, Healy, & Cameron, 2009) to estimate the intervention effects in the presence of the nested data structures. In a nested data structure, the observations are nonindependent: Children from the same classroom or school are more likely to be similar in their responses than children from a different social context. If not modeled, this nonindependence may produce inaccurate standard errors (Raudenbush & Bryk, 2002). Multilevel regression models are therefore preferable to traditional regression models because of their ability to accurately estimate the standard errors by decomposing the total variance into the various hierarchical levels of the data (Snijders & Bosker, 1999).

The data sets for Grades 2–3 and 8–9 contained three measurements; these were treated as a separate level, time points within students. We fitted four-level models to represent change over time, individual student differences, differences between classrooms, and between-school differences. We examined the gains for KiVa schools compared with control schools after controlling for baseline levels of the variable of interest, gender, age, and language of instruction at school (Finnish or Swedish). The models bear a resemblance to the models in the previous study on KiVa program effects in Grades 4–6 (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011); the model specification is described in detail in the online supplemental appendix. For the dichotomous self-reports of victimization and bullying, logistic regression analysis was used.

Because gender and age are important predictors of bullying and victimization (see, e.g., Salmivalli & Voeten, 2004; Whitney & Smith, 1993), we included them as covariates in all models. It is known that not only means but also variances of bullying-related variables may differ between boys and girls (Salmivalli & Voeten, 2004). Therefore, student-level variance was specified as a function of gender. In addition, gender and age were also entered as predictors at the classroom level (i.e., proportion of boys and average age of children in the classroom). The inclusion of these covariates enabled us to control for their effects and to investigate their potential interactions with the intervention effects. We added also the language of instruction into our models, because earlier analyses have shown that Swedish-speaking minority students may report lower levels of bullying and victimization than Finnish-speaking students (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). Furthermore, it has been previously argued that in Finland, the Swedish-speaking minority deviates positively

in some respects from the Finnish-speaking majority, for instance, in juvenile delinquency (Obstbaum, 2006).

There were several dummy-coded variables in the models. To test intervention effects at Waves 2 and 3 separately, we coded the three waves of data collection with two dummy variables: Time 2, or *T2* (Wave 2 = 1, other waves = 0), and *T3* (Wave 3 = 1, other waves = 0). In addition, gender (girls = 0, boys = 1), language of instruction (Finnish = 0, Swedish = 1), and condition (control school = 0, intervention school = 1) were entered into the models with dummy coding. Student age was centered around the average age of the students in Grade 2 for the Grades 2–3 data or in Grade 8 for the Grades 8–9 data (grand-mean centering, see Enders & Tofigi, 2007). This implies that the intercepts and the variance components estimated refer to these average ages rather than to the age of zero, which would not make sense.

Furthermore, at the classroom level, we included average age of the students in the classroom (*AgeCl*), which is highly correlated with grade level, and proportion of boys in the classroom (*BoyCl*) to test the difference between within-classroom and between-classroom regressions (see, e.g., Snijders & Bosker, 1999, pp. 27–29, and especially pp. 52–56) and to take the influence of the classroom context on bullying and victimization into account. Given our grand-mean centering and inclusion of classroom means in the models, the student-level coefficients of age and gender in Tables 4 and 5 measure the effects of age and gender within classrooms; the classroom-level coefficients for age and gender measure the extent to which these effects at the classroom level differ from those at the student level. The classroom-level coefficients therefore indicate whether the classroom average age or the gender composition

Table 4

Multilevel Modeling Results: Intervention Effects for Self-Reported Victimization and Bullying in Grades 2–3 and 8–9

Variable	Grades 2–3		Grades 8–9	
	Victimization	Bullying	Victimization	Bullying
Baseline				
Intercept	–1.38***	–3.31***	–2.37***	–3.00***
Student level				
Boy	0.48***	1.22***	0.53***	0.80***
Age	—	—	0.09	0.12*
Classroom level				
BoyCl	–0.06	—	0.50	—
AgeCl	—	—	–0.29**	—
School level				
Swedish	–0.40**	–0.84***	—	—
Intervention	–0.02	0.04	0.03	–0.13
Intervention × Boy	–0.10	—	–0.16	—
Intervention × BoyCl	0.91	—	–1.00*	—
Change by Wave 2				
T2	–0.44***	–0.24*	–0.40***	–0.26***
Student level				
Boy × T2	—	—	0.18*	—
Age × T2	—	—	0.18**	—
School level				
Intervention × T2	–0.21*	–0.34*	–0.19*	0.08
Change by Wave 3				
T3	–0.22*	–0.15	–0.54***	–0.15
Student level				
Boy × T3	–0.29*	—	0.36***	—
Classroom level				
BoyCl × T3	0.75	—	—	—
School level				
Intervention × T3	–0.49**	–0.36*	–0.04	–0.08
Intervention × Boy × T3	0.44*	—	—	—
Intervention × BoyCl × T3	–1.67*	—	—	—
Variance components				
Student level				
Baseline for girls	1.965***	6.879***	3.871***	6.414***
Baseline for boys	1.626***	2.591***	2.742***	3.015***
Classroom level				
Intercept	0.167***	0.575**	0.382***	0.432***
Boy	—	1.053**	0.601***	0.530**
School level				
Intercept	0.048*	0.014	0.088**	0.065**

Note. Estimates of covariances omitted. An em dash indicates that the estimate was not included in the model. Cl = classroom level; T2 = Time 2; T3 = Time 3.

* $p < .05$. ** $p < .01$. *** $p < .001$ (one-tailed tests for variances).

Table 5
Multilevel Modeling Results: Intervention Effects for Peer-Reported Victimization, Bullying, and Bystander Behaviors in Grades 8–9

Variable	Victimization	Bullying	Bystander behavior		
			Assisting	Reinforcing	Defending
Baseline					
Intercept	−0.09	−0.25***	−0.34***	−0.40***	0.40***
Student level					
Boy	0.19***	0.61***	0.80***	0.92***	−0.71***
Age	0.07*	0.05**	—	0.04	0.03
Classroom level					
BoyCl	−0.39**	−0.10	−0.45**	−0.30	0.26
AgeCl	0.03	0.01	—	−0.07	−0.12
School level					
Intervention	−0.03	−0.01	0.03	−0.09	0.02
Intervention × Boy	—	0.03	0.00	0.03	0.03
Intervention × Age	−0.09*	—	—	0.02	−0.01
Intervention × BoyCl	—	0.27	0.33	0.34	−0.40
Intervention × AgeCl	0.13	—	—	0.17*	0.05
Change by Wave 2					
T2	0.10***	0.03	0.05**	0.05*	−0.01
Student level					
Boy × T2	−0.04*	−0.02	−0.05*	−0.05*	0.08***
Age × T2	0.02	−0.02	—	—	—
Classroom level					
BoyCl × T2	0.32***	0.11	0.30***	0.07	0.26**
AgeCl × T2	−0.21***	—	—	—	—
School level					
Intervention × T2	−0.06**	0.02	−0.01	−0.02	−0.05*
Intervention × Boy × T2	—	−0.04	−0.03	−0.03	−0.03
Intervention × Age × T2	0.01	—	—	—	—
Intervention × BoyCl × T2	—	−0.31	−0.59***	−0.21	0.25*
Intervention × AgeCl × T2	0.13**	—	—	—	—
Change by Wave 3					
T3	0.14***	0.08***	0.09***	0.12***	−0.14***
Student level					
Boy × T3	−0.11***	−0.11***	−0.12***	−0.15***	0.18***
Age × T3	−0.02	−0.06***	—	−0.03	−0.03
Classroom level					
BoyCl × T3	—	−0.01	—	—	0.35***
AgeCl × T3	−0.15***	—	—	−0.08***	0.10*
School level					
Intervention × T3	−0.10***	0.00	−0.06*	−0.02	−0.10***
Intervention × Boy × T3	—	−0.09**	−0.10*	−0.16***	—
Intervention × Age × T3	0.08**	—	—	—	0.05
Intervention × BoyCl × T3	—	−0.36**	—	—	—
Intervention × AgeCl × T3	0.02	—	—	—	−0.11*
Variance components					
Student level					
Baseline for girls	0.694***	0.366***	0.382***	0.426***	0.696***
Baseline for boys	0.793***	0.768***	0.810***	0.702***	0.501***
T2	0.656***	0.509***	0.522***	0.568***	0.665***
T3	0.801***	0.651***	0.670***	0.725***	0.834***
Classroom level					
Intercept	0.183***	0.101***	0.141***	0.226***	0.200***
School level					
Intercept	0.024*	0.011**	0.014**	0.023**	0.033***

Note. Estimates of covariances omitted. An em dash indicates that the estimate was not included in the model.

Cl = classroom level; T2 = Time 2; T3 = Time 3.

* $p < .05$. ** $p < .01$. *** $p < .001$ (one-tailed tests for variances).

has effects on the outcome over and above the individual student's age or gender.

The models contained several interaction terms. The intervention effects were represented by the coefficients for the interaction terms Intervention × T2 and Intervention × T3, and they can be

interpreted as intervention-control differences in the average change scores by Waves 2 and 3. More specifically, the model was defined as follows: \hat{Y}_{tijk} is used to indicate time points, i is used for individual students, j is used to denote classrooms, and k to denote schools:

$$\begin{aligned}\hat{Y}_{ijk} = & b_{0ijk} + b_{1i}T2_{ijk} + b_{2i}T3_{ijk} + b_{3i}Boy_{ijk} + b_{4i}Age_{ijk} \\ & + b_{5i}BoyCl_{jk} + b_{6i}AgeCl_{jk} + b_{7i}BoyT2_{ijk} + b_{8i}BoyT3_{ijk} \\ & + b_{9i}AgeT2_{ijk} + b_{10i}AgeT3_{ijk} + b_{11i}BoyClT2_{ijk} \\ & + b_{12i}BoyClT3_{ijk} + b_{13i}AgeClT2_{ijk} + b_{14i}AgeClT3_{ijk} \\ & + b_{15i}Swedish_k + b_{16i}Intervention_k + b_{17i}InterventionT2_{ijk} \\ & + b_{18i}InterventionT3_{ijk}.\end{aligned}$$

Furthermore, we examined moderator effects of gender and age of the student, to see whether intervention effects were dependent on student or classroom characteristics. This was tested by including the terms Intervention \times Age \times T2, Intervention \times Classroom Average Age \times T2, Intervention \times Age \times T3, Intervention \times Classroom Average Age \times T3, Intervention \times Boy \times T2, Intervention \times Classroom Proportion of Boys \times T2, Intervention \times Boy \times T3, Intervention \times Classroom Proportion of Boys \times T3. We tested these interactions with multivariate Wald tests. Statistically nonsignificant interaction effects were removed from the models unless they were needed because of significant higher order interactions. To achieve model parsimony and convergence, we omitted some random effects, and we deleted all covariates with statistically nonsignificant effects from the models (for details of final models, see Tables 4 and 5). The random part of the models was kept as simple as possible and the same across dependent variables, but variance heterogeneity by gender was allowed at the student level. Random slopes of the time variables were introduced both at the classroom and school levels but were omitted in all final models because of estimation problems with some models (in these models, it appeared not to be possible to get appropriate parameter estimates when random slopes for the time variables were included). In those cases, estimates were obtained that implied correlations larger than 1.0 between slope and intercept. Parameter estimates remained practically the same whether the random slopes were in the model. So, conclusions were not affected by this simplification of the models. At the classroom level, random slopes for gender were allowed when they were statistically significant. But in the models for Table 5, these slopes for gender were removed, because in these variables there were only minor differences in classroom-level variances between boys and girls and because the removal did not affect the other results.

When statistically significant moderator effects were found, predictors were recentered to test the simple slopes (Aiken & West, 1991). Age was recentered from the average value at Grade 8 (or Grade 2) to the average value in Grade 9 (Grade 3) to test the simple slopes, both at the student and classroom levels. If a moderator effect of gender was found, *boy* was recoded at student level (boy = 0, girl = 1), and the *classroom proportion of boys* was recentered to various values ranging from low (35% boys; $M - 1$ SD), to average (50% boys), to high (65% boys; $M + 1$ SD).

Results From the Multilevel Models

We used in total seven dependent variables: self-reported bullying and victimization, peer-reported bullying and victimization, and three peer-reported bystanders' behaviors (assisting, reinforcing,

defending); the peer reports were obtained only in Grades 7–9. Because of skewness in the distributions of the continuous dependent variables, we transformed them into normal scores (Blom, 1958). The method used creates z scores corresponding to the estimated cumulative proportions. After transformation, the new distribution resembled more closely a normal distribution than the raw-score distribution did (e.g., Crocker & Algina, 2008, pp. 442–444). To achieve simplicity and brevity of presentation, we explicated only the Wave-3 results for Grades 2–3 and 8–9 in the text. The interested reader may want to compare them with the Wave-2 results included in the tables. A summary of the Grade-1 and Grade-7 results (from the posttest-only design) is provided in the online supplementary appendix.

Statistical tests. Tables 4 and 5 present the parameter estimates for the final models for each criterion variable in Grades 2–3 and 8–9, including the results for Waves 2 and 3 (T2 and T3). Unstandardized regression coefficients and variance components are reported, and for the dichotomized self-reports, logistic regression coefficients are shown (Table 4). There were residual variances at Levels 2, 3, and 4 but not at Level 1, because two dummy variables were used to represent the three time points. The tables contain only the residual variances at the student, classroom, and school levels plus random slope variances at the student (Table 5) or classroom (Table 4) levels, omitting all covariances.

Baseline equivalency between control and intervention schools. In Tables 4 and 5, the coefficients for the *intervention* variable represent the differences between control and intervention schools at baseline. In examining the descriptive statistics, we already noted that these differences were small (Table 2), and here it can be seen that they are not statistically significant, except for a few interaction effects.

Intervention effects at Wave 3: Grades 2–3 and 8–9. Intervention results concerning self-reported bullying and victimization are reported in Table 4. Compared with the control school students, second and third grade boys and girls in KiVa schools bullied less ($b = -0.36$, $p = .036$). For victimization, however, the effect depended on gender at both the student ($b = 0.44$, $p = .017$) and the classroom ($b = -1.67$, $p = .029$) levels. In other words, there were two separate interaction effects: Intervention \times Boy \times T3 and Intervention \times Classroom Proportion of Boys \times T3. Together, these interactions imply that the significant reduction of victimization associated with the intervention. Intervention \times T3, was restricted to girls in classrooms with an average proportion (50%) of boys ($b = -0.49$, $p = .001$). This reduction became even stronger when the proportion of boys increased (with 65% boys, $b = -0.74$, $p < .001$). The reduction of victimization, however, was not significant for girls in classrooms with a low proportion (35%) of boys ($b = -0.23$, $p = .179$). For boys, the reduction of victimization at Wave 3 approached statistical significance only in classrooms with 65% boys ($b = -0.30$, $p = .055$). For students in Grades 8–9, the intervention showed no statistically significant effects on self-reported bullying or victimization.

The intervention effects on peer-reported outcomes in Grades 8–9 are presented in Table 5. The intervention reduced peer-reported victimization ($b = -0.10$, $p < .001$). There was an interaction with age of student ($b = 0.08$, $p < .01$), however. Victimization decreased significantly for younger students (at or below the average for students in Grade 8), but for students who

were at the average age for Grade 9, there was hardly any effect ($b = -0.01, p = .670$).

The intervention effect on peer-reported bullying was statistically nonsignificant ($b = 0.00, p = .854$; Table 5). But because of interaction effects (i.e., Intervention \times Boy \times T3 and Intervention \times Classroom Proportion of Boys \times T3), this result applies only to girls in classrooms with an average proportion of boys. At the student level, there was a significant interaction with gender ($b = -0.09, p < .01$), and the interaction was significantly stronger at the classroom than at the individual level ($b = -0.36, p = .008$). By probing the interactions, we found that bullying was reduced for boys and the more so when the proportion of boys in the classroom was higher (35% boys, $b = -0.04, p = .237$; 50% boys, $b = -0.09, p < .001$; 65% boys, $b = -0.15, p < .001$). Bullying was not reduced for girls, but the effect approached statistical significance when a girl was in a classroom with a high proportion of boys (with 65% boys, $b = -0.07, p = .060$).

According to the Table 5 results, the intervention reduced assisting ($b = -0.06, p = .010$). This intervention effect applied to girls; for boys, the intervention effect was stronger, shown by the significant interaction with gender at the student level ($b = -0.10, p = .001$). More specifically, a significant reduction of assisting was seen for both girls ($b = -0.06, p = .010$) and boys ($b = -0.16, p = .023$).

For peer-reported reinforcing, the intervention effect was statistically nonsignificant ($b = -0.02, p = .473$; Table 5). Due to an interaction, this result, however, applies only to girls: There was an interaction effect with gender at the student level ($b = -0.16, p < .001$). Although for girls, the intervention effect at Wave 3 was not statistically significant; for boys, a larger and statistically significant reduction of reinforcing was observed ($b = -0.18, p < .001$).

The intervention effect on defending was statistically significant, but it was not in the expected direction ($b = -0.10, p < .001$; Table 5). That is, defending the victims decreased in the intervention condition. The intervention effect appeared the same for boys and girls but differed by age. The negative effect on defending did not depend on the age of the student ($b = 0.05, p = .292$) but increased with the average age in the classroom ($b = -0.11, p = .041$). There was no significant effect for older students in a classroom with primarily younger classmates (more specifically, for students whose age was equal to the Grade-9 average but who were in classrooms with average age equal to the Grade-8 average [$b = -0.06, p = .219$]). For students in Grade-9 classrooms, the negative effect on defending became larger ($b = -0.16, p < .001$ and $b = -0.21, p < .001$, for average student ages at Grades 8 and 9, respectively).

Effect sizes. We calculated effect sizes for the intervention at Waves 2 and 3: model-based odds ratios (ORs) for the dichotomous self-reports and Cohen's d s for the continuous variables (Tables 6 and 7). The calculations were done assuming an average proportion of boys in classroom (about 50%). The odds ratios were converted to represent the odds of being a bully or a victim in a control school compared with the respective odds in an intervention school. For Cohen's d s, an effect size with a positive sign stands here for a positive (i.e., desired) intervention effect.

The Wave-3 results show that in Grades 2–3, the odds of being a victim or a bully were approximately 1.5 times larger in the control schools than in the intervention schools. The only exception was for boys in Grades 2–3, for whom there was no interven-

Table 6
Odds Ratios and 95% Confidence Intervals for the Intervention Effects in Grades 2–3 and 8–9

Variable	Wave 2 odds ratio [95% CI]	Wave 3 odds ratio [95% CI]
Grades 2–3		
Victimization	1.23 [1.04, 1.42]	—
Girls	—	1.63 [1.34, 1.91]
Boys	—	1.04 [0.79, 1.30]
Bullying	1.41 [1.07, 1.75]	1.43 [1.10, 1.77]
Grades 8–9		
Victimization	1.21 [1.04, 1.38]	1.04 [0.86, 1.22]
Bullying	0.93 [0.73, 1.12]	1.08 [0.88, 1.28]

Note. A confidence interval that does not include 1.00 implies that $p < .05$. An em dash indicates that the effect size was not calculated. CI = confidence interval.

tion effect on victimization ($OR = 1.04$). For Grades 8–9, the results were not statistically significant.

In Grades 8–9, the largest effects on peer reports were obtained at Wave 3 for boys' reinforcing ($d = 0.19$), assisting ($d = 0.18$), and bullying ($d = 0.11$), and for victimization among both boys and girls in Grade 8 ($d = 0.10$). Many effect sizes for the peer-reported variables in Grades 8 and 9 depended on the proportion of boys in the classroom. Typically, positive intervention effects increased when the proportion of boys was higher. All other effect sizes were small, and several of them rather close to zero.

To make the results of the KiVa program comparable with previous studies (Farrington & Ttofi, 2010), we estimated how much the entire KiVa program reduced the odds for bullying and victimization and the prevalence of these problems. To this end, we used the whole KiVa project sample comprising of Grades 1–9 students from the two phases of the randomized controlled trial evaluation of the program. We included into the calculations those students who had gained parental permission to participate and who had responded both at pretest and posttest (except for Grades 1 and 7 for which only posttest was included). The total sample size was 24,138 for victimization and 24,002 for bullying (with response rates of 70% and 69%). We categorized the students into victims, bullies, and noninvolved children by dichotomizing the self-reported bullying and victimization, as described previously. We calculated odds ratios for victimization and bullying, controlling for the pretest differences (except for Grades 1 and 7, where it was not possible). Next, average weighted means were calculated across the grade levels, and the standard errors were corrected for clustering at the school level by multiplying them with the design effect (based on the ICCs and the average school sizes in our sample; for the formulas, see Farrington & Ttofi, 2010). For victimization, the odds ratios were 1.33, 1.53, and 1.13, and for bullying 1.50, 1.41, and 1.21 in Grades 1–3, 4–6, and 7–9, respectively. The effect sizes were larger in primary school (Grades 1–6) than in secondary school (Grades 7–9). The average weighted odds ratios across all grade levels were for victimization 1.28 with 95% confidence interval (CI) [1.17, 1.40] and for bullying 1.30 with 95% CI [1.15, 1.48]. Therefore, the odds of being a victim or being a bully were about 1.3 times higher for a control-school student than for a student in an intervention school. This corresponds to a reduction of about 20% in the prevalence of bullying and victimization.

Table 7
Cohen's *ds* for the Intervention Effects in Grades 8–9

	Wave 2	Wave 3
Peer-reported victimization		
Grade 8	0.06	0.10
Grade 9	–0.08	0.01
Peer-reported bullying		
Boys	0.02	0.11
Girls	–0.03	0.00
Peer-reported assisting		
Boys	0.04	0.18
Girls	0.01	0.06
Peer-reported reinforcing		
Boys	0.05	0.19
Girls	0.02	0.02
Peer-reported defending		
Boys	–0.08	—
Girls	–0.05	—
Grade 8	—	–0.11
Grade 9	—	–0.17

Note. Effect sizes were computed as gain for intervention group minus gain for control group. Cohen's *d* was calculated as the adjusted group mean difference divided by unadjusted pooled within-group standard deviation:

$$d = \frac{b}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2}}}$$

where *b* is the coefficient for the intervention's effect, which represents the group mean difference adjusted for student- and school-level covariates (Intervention × T2 or Intervention × T3); *n*₁ and *n*₂ are the student-level sample sizes; and *SD*₁ and *SD*₂ are the student-level unadjusted pretest standard deviations for the intervention group and the control group, respectively. The sign of *b* was determined such that a positive *d* always signifies a positive (i.e., desired) intervention effect. All the results are provided in the table regardless of their statistical significance. An em dash indicates that the effect size was not calculated.

Discussion

The present study examined the effectiveness of the KiVa program for Grades 1–3 and 7–9, and it thereby complements the previous findings for Grades 4–6 (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). The results indicate that KiVa is effective in reducing bullying and victimization not only in Grades 4–6 but in Grades 1–3 as well. In Grades 8–9, there were significant positive effects on four of the five peer-reported outcomes, but these effects depended on student and classroom characteristics. In contrast to Grades 4–6, no significant positive effects were found on any of the other outcomes.¹ As a whole, the intervention effects on bullying and victimization appeared larger and more consistent in elementary than in lower secondary schools. The current study thus makes a unique contribution to the literature (a) by providing new knowledge about the effectiveness of the KiVa antibullying intervention program and (b) by supplying evidence about the effectiveness of the program for students and classrooms varying in age and gender.

A comparison of the effect sizes across the present and the previous study (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011), provides a somewhat complicated pattern of results. Nevertheless, on the basis of odds ratios calculated in the same way for the whole KiVa sample, it seems that the intervention effects on

self-reported victimization and bullying are generally larger in Grades 1–6 than in Grades 7–9. Compared with previous studies, the overall effect sizes in the present study for Grades 1–9 (*ORs* 1.28 and 1.30 for victimization and bullying, respectively) correspond well to the results of a recent meta-analysis (Farrington & Ttofi, 2010), which showed that the average odds ratio for anti-bullying intervention programs (excluding the KiVa program) was 1.29 (95% CI [1.17, 1.41]) for victimization and 1.36 (95% CI [1.26, 1.47]) for bullying (David P. Farrington, personal communication, March 8, 2010). Furthermore, the KiVa program effects are larger than the average effects for studies with randomized design for victimization (*OR* = 1.17, 95% CI [1.00, 1.37]) and for bullying (*OR* = 1.10, 95% CI [0.97, 1.26]; Farrington, 2010).

With regard to other outcomes measured only in Grades 4–9, the intervention effects were larger and more consistent in Grades 4–6 than in Grades 7–9. In both samples, the intervention effects on peer-reported bullying, victimization, assisting, and reinforcing were at least equally large compared with the effects on self-reported bullying and victimization.

In Grades 8 and 9, positive and significant effects were found on four of the five peer-reported role scales (victimization, bullying, assisting, and reinforcing). But the size of the intervention effects depended on gender and sometimes also on age. In several instances, there were stronger effects for boys and in classrooms with a high proportion of boys. No statistically significant effects were found in Grade 7 for the peer-reported outcomes, and most of the effect sizes were close to zero. These results are less dependable because of the lacking pretest, however. We had no specific prior hypotheses about the moderating effects of gender or age, which makes the analyses somewhat explorative, and the results should be replicated in further studies. The stronger positive effects for boys may be a consequence of boys' high scores that make them suitable targets for the intervention to reduce bullying, assisting, and reinforcing. It is an interesting result that the effects on individual students were largest in classrooms with a high proportion of boys. Perhaps in such classrooms there is the largest potential for improvement because a large concentration of boys may lead to an increase of problematic behaviors; this increase may be counteracted by the intervention.

It has been proposed that as children turn into adolescents, their social intelligence increases, and this may cause (a) a decrease in physical and verbal aggression and (b) an increase in indirect aggression (e.g., Björkqvist, Lagerspetz, & Kaukiainen, 1992; Björkqvist, Österman, & Kaukiainen, 1992). The KiVa program can be less effective in reducing the indirect forms of bullying and victimization, which may partially account for the weaker effects in the lower secondary schools compared with the primary schools. It is also possible that as students age, it becomes increasingly difficult to influence the bullying-related classroom norms (i.e., that a mediation effect of norms is moderated by age). Further studies are needed to investigate these possibilities empirically.

¹ We examined the intervention effects also on antibullying attitudes, empathy toward victims, self-efficacy for defending and wellbeing at school in Grades 7–9 (cf. Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011, for a definition of these variables). The effect sizes for the intervention effects on these outcomes were in Grades 7–9 practically zero; to conserve space, these analyses were not reported in the present article.

The effect sizes discussed previously can be considered small or moderate. The significant interactions imply, however, that the intervention effects may differ for different types of students. If that is true, then average effect sizes for victimization and bullying (whether self-reports or peer reports) and for reinforcing and assisting may look somewhat modest because of that. This is apparent in the interaction effects with student and classroom characteristics found in the data from Grades 8–9. The magnitude of the effect sizes may also be related to the high proportion of (consistently) noninvolved students. Students who during the school year were never involved in bullying and who were never victimized cannot show an intervention effect. Only when the proportion of such students is lower in intervention schools than in control schools can we have an intervention effect. Intervention effects in the form of reducing negative behaviors can only be found for those students who were involved in bullying or victimization. This may also be related to the finding that intervention effects for some variables were stronger for boys than girls and stronger in classrooms with higher rather than lower proportions of boys. Furthermore, the difference in results for self-reported and peer-reported bullying and victimization may in part be explained by the fact that there was a decrease in the control condition for the self-reports but not for the peer reports. If the problematic behaviors decrease to some extent even without an intervention (or with “treatment as usual,” as Finnish schools are obliged by law to counteract bullying), detecting the intervention effects becomes more difficult.

For defending the victims, a significant effect in the wrong direction was found: Compared with their peers in control schools, the students in intervention schools on average defended the victims less. This is a surprising finding, because it suggests that turning adolescents into defenders of victims is more difficult in Grades 7–9 than in Grades 4–6 (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). These results are in contradiction with a recent meta-analysis by Polanin, Espelage, and Pigott (2012) in which the average effect size for the bystander intervention was larger for older (Grades 9–12; $g = .43$) than for younger (Grades 3–8; $g = .14$) students. If the KiVa results were indeed replicated, it would be an important task to investigate in detail the processes involved. A partial explanation may be that KiVa did not succeed in Grades 7–9 in increasing students’ antibullying attitudes, empathy toward victims, or self-efficacy for defending: These are characteristics that have been found to predict defending behavior (Pöyhönen et al., 2010; Salmivalli & Voeten, 2004).

Limitations

The fact that there were only posttest data for students in Grades 1 and 7 makes it impossible to control for potential preexisting differences between intervention and control conditions, and it weakens the evidence for these grade levels. Furthermore, for students in Grades 1–3, the outcome variables included only self-reports of bullying and victimization. These young students’ global self-reports correlated with respective questions about different forms of bullying and victimization equally well as for older students in Grades 4–9. This indicates that although only self-reports were gathered in Grades 1–3, they were as such valid measures of the phenomena under study. Another limitation is that we did not investigate the effectiveness of KiVa on different forms

of victimization and bullying. It remains an important topic for future studies to find out whether the relative impact of the KiVa program varies by the specific form of victimization or bullying (e.g., physical, verbal, or indirect) measured at various grade levels. Yet another important limitation is that the results were assessed solely by questionnaire data. It is possible that the less favorable results in the higher grades may partially be a consequence of the measurement method: There were signs (i.e., implausible or impossible responses) suggesting that the students were not always answering the questions sincerely. Finally, student surveys were administered by teachers. When students answer questions about undesirable behaviors in the presence of their teacher, they may be influenced to answer the way they think the teacher would like. We sought to prevent this by giving teachers detailed instructions on how to act in the survey administration, and the students were told that their answers remain confidential.

In the field trials of intervention programs, several threats to the internal validity of conclusions must be dealt with (e.g., Shadish, Cook, & Campbell, 2002). Most of the alternative explanations for the results obtained in the present study can be considered rather implausible, because we used (a) random assignment of schools into intervention and control conditions and (b) FIML to deal with missing data. Although it was not possible to actually test the MAR assumption of FIML, methodologists have demonstrated that in many realistic cases, failing to take into account a cause or correlate of missingness has only a minor impact on estimates and standard errors (Collins et al., 2001).

Some of the intervention schools ($n = 31$) participating in the present study were randomized into a control condition for Grades 4–6, and during the school year 2007–2008, they were on a waiting list to receive the KiVa program. These schools were promised that they would be included in the intervention condition during the following school year, because otherwise some of them might have easily dropped out of the study. This waiting list may be considered a limitation of our present randomization procedure. But it should be kept in mind that (a) all the schools were originally randomized into intervention and control conditions, (b) they belonged to the same pool of volunteers, and (c) we might have ended up with the present samples even without the waiting list element. The waiting list procedure is unlikely to have had any noticeable effect on the results. The only obvious consequence for the measurements is that a subgroup of Grade 7 students had previously answered the study questionnaires as Grade 6 students, but these students were not included in the main analyses, which involved only Grades 8–9.

With regard to external validity, it can be noted that we had a diverse sample of schools from all over the mainland Finland, including both Finnish- and Swedish-language schools. On the other hand, all our schools were volunteering to take part into the study, and during the study, there was some attrition, with a larger proportion of more problematic students dropping out. Such attrition limits the generalizability of our results to some extent, and further studies are needed to investigate the effectiveness of the KiVa program when it is disseminated widely to a larger sample of schools. Actually, one such study has already been published (Kärnä, Voeten, Little, Poskiparta, Alanen, et al., 2011). There we found on the basis of a sample consisting of 888 schools that the KiVa program also was effective when disseminated broadly throughout the country. The effects were somewhat smaller (the

ORs for self-reported bullying and victimization equaled about 1.2), but the pattern of effects was similar to the present study: larger effects in elementary schools compared with lower secondary schools. Another possible limitation is that some cultural specificities of the Finnish context or school system have contributed to the differential effectiveness of the KiVa program. There are research projects in progress in both the Netherlands and the United States to investigate the effectiveness of the KiVa program; the results from these evaluations will provide some idea about the relevance of the school system for the effectiveness of the KiVa program.

Implications

Considered as a whole, the results from the present and the previous study (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011) support the view that after 9 months of implementation, the KiVa program is effective in primary school (Grades 1–6), whereas the positive effects in secondary school (Grades 7–9) are more modest and depend especially on the gender of the student. These findings are in contrast with the conclusion of Farrington and Ttofi (2010) that the effects of antibullying programs increase steadily as a function of age. They even suggested that “[antibullying] programs should be targeted on children aged 11 or older rather than on younger children” (Farrington & Ttofi, 2010, p. 72). According to our findings, even students in Grades 1–3 can benefit significantly from antibullying programs, whereas it may be much more difficult to reduce bullying and victimization among secondary school students. Actually, the results from the present study are in accordance with views of P. K. Smith (2010), who argued that the interventions are less effective in secondary schools than in primary schools. He proposed several explanations for the modest results, such as developmental changes related to adolescence (e.g., changes in peer relations), difficulty of change in large organizations such as secondary schools, and differences in teacher roles between primary and secondary schools. These are all possible explanations for the differences in effectiveness of the primary- and secondary-school versions of the KiVa program.

Future Directions

The ultimate aim of the Finnish Ministry of Education and Culture was to develop a research-based antibullying program that could be used in all Finnish elementary and lower secondary schools. The dissemination of KiVa to Finnish schools started in 2009, and after the first 3 years, 90% of all schools in the country have joined in. As the students in the participating schools will answer questionnaires every spring, this will create a unique opportunity to investigate the long-term effects of KiVa on bullying and victimization.

After the present and previous studies (e.g., Kärnä, Voeten, Little, Poskiparta, Alanen, et al., 2011; Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011) on the main effects of KiVa, several important questions remain for future research. For instance, it is important to try to understand why the effects of the KiVa program seem to be larger for primary schools than for secondary schools. This requires investigation of mediators and moderators of program effects in the different age groups at the multiple systemic levels of student, classroom, and school. The

degree of fidelity of program implementation may, to some extent, explain variation in the intervention outcomes (e.g., Olweus & Alsaker, 1991; Salmivalli et al., 2005; Whitney, Rivers, Smith, & Sharp, 1994). Research on the association between implementation and intervention results will give some idea of how much the intervention results can be improved by providing support for schools in program implementation. Finally, investigating the predictors of implementation (Kallestad & Olweus, 2003) will provide information about what kind of schools need additional resources for high-quality implementation of the KiVa program.

References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- Bandura, A. (1989). Social cognitive theory. In R. Vasta (Ed.), *Annals of child development: Six theories of child development* (pp. 1–60). Greenwich, CT: JAI Press.
- Björkqvist, K., Lagerspetz, K. M. J., & Kaukiainen, A. (1992). Do girls manipulate and boys fight? *Aggressive Behavior*, 18, 117–127.
- Björkqvist, K., Österman, K., & Kaukiainen, A. (1992). The development of direct and indirect aggressive strategies in males and females. In K. Björkqvist & P. Niemelä (Eds.), *Of mice and women: Aspects of female aggression* (pp. 51–64). San Diego, CA: Academic Press.
- Blom, G. (1958). *Statistical estimates and transformed beta variables*. New York, NY: Wiley.
- Caravita, S., DiBlasio, P., & Salmivalli, C. (2009). Unique and interactive effects of empathy and social status on involvement in bullying. *Social Development*, 18, 140–163. doi:10.1111/j.1467-9507.2008.00465.x
- Cillessen, A. H. N., & Mayeux, L. (2004). From censure to reinforcement: Developmental changes in the association between aggression and social status. *Child Development*, 75, 147–163. doi:10.1111/j.1467-8624.2004.00660.x
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330–351. doi:10.1037/1082-989X.6.4.330
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12, 121–138. doi:10.1037/1082-989X.12.2.121
- Eslea, M., & Smith, P. K. (1998). The long-term effectiveness of anti-bullying work in primary schools. *Educational Research*, 40, 203–218. doi:10.1080/0013188980400208
- Farrington, D. P., & Ttofi, M. M. (2010). School-based programs to reduce bullying and victimization. *Campbell Collaboration Library of Systematic Reviews*, 6. Retrieved from <http://www.campbellcollaboration.org/library.pl>
- Ferguson, C., San Miguel, C., Kilburn, J., & Sanchez, P. (2007). The effectiveness of school-based anti-bullying programs: A meta-analytic review. *Criminal Justice Review*, 32, 401–414. doi:10.1177/0734016807311712
- Hanewinkel, R. (2004). Prevention of bullying in German schools: An evaluation of an anti-bullying approach. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 81–98). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511584466.006
- Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, 45, 1195–1199. doi:10.1037/a0015665
- Juvonen, J., & Galván, A. (2008). Peer influence in involuntary social

- groups: Lessons from research on bullying. In M. J. Prinstein & K. A. Dodge (Eds.), *Understanding peer influence in children and adolescents* (pp. 225–244). New York, NY: Guilford Press.
- Juvonen, J., Graham, S., & Schuster, M. A. (2003). Bullying among young adolescents: The strong, the weak, and the troubled. *Pediatrics*, 112, 1231–1237. doi:10.1542/peds.112.6.1231
- Kallestad, J. H., & Olweus, D. (2003). Predicting teachers' and schools' implementation of the Olweus Bullying Prevention Program: A multi-level study. *Prevention & Treatment*, 6, 3–21. doi:10.1037/1522-3736.6.1.621a
- Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Alanen, E., & Salmivalli, C. (2011). Going to scale: A nonrandomized nationwide trial of the KiVa antibullying program for Grades 1–9. *Journal of Consulting and Clinical Psychology*, 79, 796–805. doi:10.1037/a0025740
- Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011). A large-scale evaluation of the KiVa antibullying program: Grades 4–6. *Child Development*, 82, 311–330. doi:10.1111/j.1467-8624.2010.01557.x
- Kärnä, A., Voeten, M., Poskiparta, E., & Salmivalli, C. (2010). Vulnerable children in varying classroom contexts: Bystanders' behaviors moderate the effects of risk factors on victimization. *Merrill-Palmer Quarterly*, 56, 261–282. doi:10.1353/mpq.0.0052
- Menesini, E., Codecasa, E., Benelli, B., & Cowie, H. (2003). Enhancing children's responsibility to take action against bullying: Evaluation of a befriending intervention in Italian middle schools. *Aggressive Behavior*, 29, 1–14. doi:10.1002/ab.80012
- Merrell, K., Gueldner, B., Ross, S., & Isava, D. (2008). How effective are school bullying intervention programs? A meta-analysis of intervention research. *School Psychology Quarterly*, 23, 26–42. doi:10.1037/1045-3830.23.1.26
- Obstbaum, Y. (2006). *Brottslighet bland finskspråkiga och svenskspråkiga ungdomar* [Crime among Finnish-speaking and Swedish-speaking young people] (Oikeuspoliittisen tutkimuslaitoksen tutkimustiedonantoja 69) [National Research Institute of Legal Policy Research Communications 69]. Helsinki, Finland: National Research Institute of Legal Policy.
- Olweus, D. (1996). *The Revised Olweus Bully/Victim Questionnaire*. Bergen, Norway: University of Bergen, Research Center for Health Promotion (HEMIL Center).
- Olweus, D. (2004). The Olweus Bullying Prevention Programme: Design and implementation issues and a new national initiative in Norway. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 13–36). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511584466.003
- Olweus, D. (2005). *New positive results with the Olweus Bullying Prevention Program in 37 Oslo schools*. Unpublished report, University of Bergen, Research Center for Health Promotion (HEMIL Center), Bergen, Norway.
- Olweus, D., & Alsaker, F. (1991). Assessing change in a cohort-longitudinal study with hierarchical data. In D. Magnusson, L. Bergman, G. Rudinger, & B. Törestad (Eds.), *Problems and methods in longitudinal research: Stability and change* (pp. 107–132). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511663260.008
- Pikas, A. (1989). The common concern method for the treatment of mobbing. In E. Roland & E. Munthe (Eds.), *Bullying: An International Perspective* (pp. 91–104). London, England: David Fulton.
- Pitts, J., & Smith, P. K. (1995). *Preventing school bullying*. London, England: Home Office.
- Polanin, J. R., Espelage, D. L., & Pigott, T. D. (2012). A meta-analysis of school-based bullying prevention programs' effects on bystander intervention behavior. *School Psychology Review*, 41, 47–65.
- Pöyhönen, V., Juvonen, J., & Salmivalli, C. (2010). What does it take to stand up for the victim of bullying? The interplay between personal and social factors. *Merrill-Palmer Quarterly*, 56, 143–163. doi:10.1353/mpq.0.0046
- Pöyhönen, V., & Salmivalli, C. (2008). New directions in research and practice addressing bullying: Focus on defending behavior. In D. Pepler & W. Craig (Eds.), *An international perspective on understanding and addressing bullying* (PREVNet Publication Series, 1, pp. 26–43). Bloomington, IN: AuthorHouse.
- Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2009). *MLwiN (Version 2.11) [Computer software]*. Bristol, England: University of Bristol, Centre for Multilevel Modelling.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Robinson, G., & Maines, B. (1997). *Crying for help: The no blame approach to bullying*. Bristol, England: Lucky Duck.
- Rodkin, P. C., Farmer, T. W., Pearl, R., & Van Acker, R. (2000). Heterogeneity of popular boys: Antisocial and prosocial configurations. *Developmental Psychology*, 36, 14–24. doi:10.1037/0012-1649.36.1.14
- Sainio, M., Kaukiainen, A., Willför-Nyman, U., Annevirta, T., Pöyhönen, V., & Salmivalli, C. (2009). *KiVa: Teacher's guide, Unit 3 (Research into Practice Publication Series, No. 4)*. Turku, Finland: University of Turku, Psychology Department.
- Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior*, 15, 112–120. doi:10.1016/j.avb.2009.08.007
- Salmivalli, C., Garandeau, C., & Veenstra, R. (2012). KiVa anti-bullying program: Implications for school adjustment. In A. M. Ryan & G. W. Ladd (Eds.), *Peer relationships and adjustment at school* (pp. 279–305). Charlotte, NC: Information Age.
- Salmivalli, C., Kärnä, A., & Poskiparta, E. (2010a). Development, evaluation, and diffusion of a national anti-bullying program KiVa. In B. Doll, W. Pfohl, & J. Yoon (Eds.), *Handbook of youth prevention science* (pp. 238–252). New York, NY: Routledge.
- Salmivalli, C., Kärnä, A., & Poskiparta, E. (2010b). From peer putdowns to peer support: A theoretical model and how it translated into a national anti-bullying program. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 441–454). New York, NY: Routledge.
- Salmivalli, C., Kärnä, A., & Poskiparta, E. (2011). Counteracting bullying in Finland: The KiVa program and its effects on different forms of being bullied. *International Journal of Behavioral Development*, 35, 405–411. doi:10.1177/0165025411407457
- Salmivalli, C., Kaukiainen, A., & Voeten, M. (2005). Anti-bullying intervention: Implementation and outcome. *British Journal of Educational Psychology*, 75, 465–487. doi:10.1348/000709905X26011
- Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior*, 22, 1–15. doi:10.1002/(SICI)1098-2337(1996)22:1<1::AID-AB1>3.0.CO;2-T
- Salmivalli, C., & Peets, K. (2008). Bullies, victims, and bully-victim relationships. In K. Rubin, W. Bukowski, & B. Laursen (Eds.), *Handbook of peer interactions, relationships, and groups* (pp. 322–340). New York, NY: Guilford Press.
- Salmivalli, C., Poskiparta, E., Tikka, A., & Pöyhönen, V. (2009). *KiVa: Teacher's guide, Unit 1 (Research into Practice Publication Series, No. 2)*. Turku, Finland: University of Turku, Department of Psychology.
- Salmivalli, C., & Voeten, M. (2004). Connections between attitudes, group norms, and behaviors associated with bullying in schools. *International Journal of Behavioral Development*, 28, 246–258. doi:10.1080/01650250344000488
- Salmivalli, C., Voeten, M., & Poskiparta, E. (2011). Bystanders matter: Associations between defending, reinforcing, and the frequency of bullying in classrooms. *Journal of Clinical Child and Adolescent Psychology*, 40, 668–676. doi:10.1080/15374416.2011.597090

- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. doi:10.1037/1082-989X.7.2.147
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton–Mifflin.
- Smith, J. D., Schneider, B. H., Smith, P. K., & Ananiadou, K. (2004). The effectiveness of whole-school anti-bullying programs: A synthesis of evaluation research. *School Psychology Review*, 33, 547–560.
- Smith, P. K. (2010). Bullying in primary and secondary schools: Psychological and organizational comparisons. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 137–150). New York, NY: Guilford Press.
- Smith, P. K., & Sharp, S. (Eds.). (1994). *School bullying: Insights and perspectives*. New York, NY: Routledge. doi: 10.4324/9780203425497
- Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage.
- Solberg, M. E., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior*, 29, 239–268. doi:10.1002/ab.10047
- Stevens, V., De Bourdeaudhuij, I., & Van Oost, P. (2000). Bullying in Flemish schools: An evaluation of anti-bullying intervention in primary and secondary schools. *British Journal of Educational Psychology*, 70, 195–210. doi:10.1348/000709900158056
- Whitney, I., Rivers, I., Smith, P. K., & Sharp, S. (1994). The Sheffield Project: Methodology and findings. In P. K. Smith & S. Sharp (Eds.), *School bullying: Insights and perspectives* (pp. 20–56). London, England: Routledge.
- Whitney, I., & Smith, P. K. (1993). A survey of the nature and extent of bullying in junior/middle and secondary schools, *Educational Research*, 35, 3–25. doi:10.1080/0013188930350101
- Williford, A., Boulton, A., Noland, B., Kärnä, A., Little, T. D., & Salmivalli, C. (2012). Effects of the KiVa Anti-Bullying Program on adolescents' perception of peers, depression, and anxiety. *Journal of Abnormal Child Psychology*, 40, 289–300. doi:10.1007/s10802-011-9551-1

Received November 22, 2011

Revision received August 9, 2012

Accepted August 28, 2012 ■

Correction to Kärnä et al. (2012)

The article “Effectiveness of the KiVa Antibullying Program: Grades 1–3 and 7–9,” by Antti Kärnä, Marinus Voeten, Todd D. Little, Erkki Alanen, Elisa Poskiparta, and Christina Salmivalli (*Journal of Educational Psychology*, Advance online publication. October 22, 2012. doi: 10.1037/a0030417) omitted some wording in the text. The sentence in the first paragraph below Table 5 beginning with, “More specifically, the model . . .” should have read “More specifically, the model was defined as follows: \hat{Y}_{ijk} is the predicted value, t is used to indicate time points, i is used for individual students, j is used to denote classrooms, and k to denote schools.”

DOI: 10.1037/a0031120

Early Adolescent Depression Symptoms and School Dropout: Mediating Processes Involving Self-Reported Academic Competence and Achievement

Cintia V. Quiroga

University of Ottawa and Children's Hospital of Eastern
Ontario, Ottawa, Canada

Michel Janosz and Sherri Bisset

University of Montreal

Alexandre J. S. Morin

University of Western Sydney

Research on adolescent well-being has shown that students with depression have an increased risk of facing academic failure, yet few studies have looked at the implications of adolescent depression in the process of school dropout. This study examined mediation processes linking depression symptoms, self-perceived academic competence, and self-reported achievement in 7th grade to dropping out of school in later adolescence. We followed 493 (228 girls and 265 boys) French-speaking adolescents from low-socioeconomic-status secondary schools in Montreal (Quebec, Canada) for 6 years. Almost 34% of participants dropped out of school during this period. Findings indicated that self-reported depression symptoms in 7th grade increased the risk of dropping out of school in later adolescence. Structural equation modeling revealed that the predictive relationship between depression symptoms and school dropout was mediated by self-perceptions of academic competence. Current findings provide support for self-perceptions of competence as mediational processes in the relationship between adolescent depression symptoms and early school leaving.

Keywords: depression, school dropout, academic competence, achievement, mediation process

A troubling number of adolescents showing serious emotional distress and depression symptoms are at risk for school failure and dropout (Quiroga, Janosz, Lyons, & Morin, 2012; Thompson, Moody, & Eggert, 1994; Wagner, Kutash, Duchnowski, Epstein, & Sumi, 2005). Despite this, there is much we do not understand about the role of depression in dropping out of school. School dropout is a complex long-term process involving multiple environmental, psychological, and academic factors (Rumberger, 2011). Depression can contribute to this process in different ways and interfere with adolescent development by impairing social, cognitive, and academic functioning (Kovacs & Goldstone, 1991;

Nolen-Hoeksema, Girgus, & Seligman, 1992). Yet the research examining the educational outcomes of adolescent depression is sparse and has shown inconsistent findings (Fergusson & Woodward, 2002; Vander Stoep, Weiss, McKnight, Beresford, & Cohen, 2002), leaving many unanswered questions about this issue. One particular matter concerns the unraveling of the processes through which depression symptoms in adolescence might become linked to school leaving.

These processes possibly involve the relationship between depression, academic competence, and achievement. Studies have shown that depression symptoms are negatively associated with self-reported academic competence and achievement and that these factors can undermine school success (Birmaher et al., 2004; Hishinuma et al., 2001; Roeser, Eccles, & Sameroff, 2000). Yet the dynamics implicating depression, self-competence, and achievement have not been examined in relation to school dropout. Theoretical models integrating mental health issues and school success suggest that depression in adolescence can compromise schooling through self-regulation processes such as student sense of competence and mastery (Roeser & Eccles, 2000; Rudolph, 2004). Hence, the objective of this research is to test a model examining the relationship between depression and school dropout that includes student self-perceptions of academic competence and self-reported achievement as mediating processes.

Depression and School Dropout

Although cross-sectional studies have shown that depressed youths are more likely to interrupt their schooling (Asarnow et al., 2005; Reinherz, Frost, & Pakiz, 1991), longitudinal research link-

This article was published Online First February 18, 2013.

Cintia V. Quiroga, School of Psychology, University of Ottawa, Ottawa, Canada, and Children's Hospital of Eastern Ontario, Ottawa, Canada; Michel Janosz and Sherri Bisset, School of Psychoeducation, University of Montreal Public Health Research Institute, and School Environment Research Group, University of Montreal, Quebec, Canada; Alexandre J. S. Morin, Centre for Positive Psychology and Education, University of Western Sydney, Sydney, Australia.

This research was supported in part by the Social Sciences and Humanities Research Council of Canada and by Fonds Québécois de la Recherche sur la Société et la Culture Grant 2006-SE-103697. This study was conducted as part of Cintia V. Quiroga's doctoral thesis at the Department of Psychology and School Environment Research Group, University of Montreal, Quebec, Canada.

Correspondence concerning this article should be addressed to Cintia V. Quiroga, Children's Hospital of Eastern Ontario, 401 Smyth Road, Room R1114, Ottawa, Ontario K1L 8L1, Canada. E-mail: cquiroga@uottawa.ca

ing depression to dropout has reported contradictory results. Namely, whereas Vander Stoep et al. (2002) showed that adolescents with depression were less likely to complete school in young adulthood, other studies found no effect of depression after adjusting for risk factors of school dropout (such as family, individual, and social experience; Fergusson & Woodward, 2002; Miech, Caspi, Moffitt, Wright, & Silva, 1999). It remains unclear whether the relationship between depression and dropout reflects the spurious effect of adverse life experiences as suggested by Fergusson and Woodward (2002), or whether depression contributes indirectly to the dropout process. It may be that the effect of depression is actually mediated by other risk factors of school dropout. In that case, it is essential to identify risk factors that link depression and dropout to understand the needs of students struggling with depressive symptoms.

Depression, Self-Reported Academic Competence, and Achievement

One of the key factors linking depression to school dropout might be student self-perception of academic competence. Studies have shown that depression and other internalizing problems are negatively associated with academic competence beliefs (Muris, Schouten, Meesters, & Gijssbers, 2003; Roeser, Strobel, & Quihuis, 2002). These symptoms can lead to the deterioration of student self-perceptions in adolescence (Cole, Martin, Peeke, Seroczynski, & Fier, 1999; Cole, Martin, & Powers, 1997; Roeser et al., 2000). Small but consistent covariations between depression symptoms, lower competence, and achievement have also been reported (Hishinuma et al., 2001; Roeser, Eccles, & Sameroff, 1998). It seems that depressed adolescents have more problematic patterns of academic functioning and achieve inferior grades than other students as reflected by lower grade point average (GPA) and self- and parent-reported academic performance (Aluja & Blanch, 2002; Birmaher et al., 2004; Puig-Antich et al., 1993; Shahar et al., 2006; Wiest, Wong, & Kreil, 1998). Yet, other studies have found no such relation (Fleming et al., 2005; Hamilton, Asarnow, & Thompson, 1997; Nurmi, Onatsu, & Haavisto, 1995). It has been suggested that the relationship between depression and academic achievement may reflect primarily lower self-perceptions of competence rather than actual academic competency levels (Aluja & Blanch, 2002; Roeser et al., 1998). Overall, although it may be unclear whether depression affects student achievement *per se*, findings highlight that youth facing emotional problems tend to hold more pessimistic self-views about school success.

Despite this evidence, few theoretical models integrating mental health issues and academic experience have been proposed, with the exception of Roeser's (Roeser & Eccles, 2000) and Rudolph's (2004) work. According to these theoretical models, depression can both compromise school success and emerge from academic difficulties. To explain how depression may affect schooling, both models "emphasize self-regulation processes as links between depression and school adjustment" (Rudolph, 2004, p. 36) that include low self-perceptions of academic competence. Roeser argued that feelings of depression can undermine school success by activating related negative motivational beliefs (Roeser & Eccles, 2000). This is important because students who believe in their ability to control school outcomes successfully are more likely to deploy the necessary efforts for achieving their goals, so positive

self-perceptions of academic competence translate into higher expectations of success (Eccles & Wigfield, 2002; Malmberg, Wanner, & Little, 2008; Skinner, 1995; Skinner, Wellborn, & Connell, 1990). Depressed individuals tend however to have more pessimistic views about themselves and future events and of their ability to successfully influence or change their outcomes (Abramson, Metalsky, & Alloy, 1989; Hankin, Abramson, & Siler, 2001). This is particularly problematic, as perceived competence relates directly to specific cognitive and behavioral aspects of student self-regulation. So students with low self-perceived competence also tend to report reduced attention, activity value, monitoring of work time, and task persistence and more achievement-related helpless behavior (Bouffard-Bouchard, Parent, & Larivée, 1991; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Nolen-Hoeksema et al., 1992; Perels, Gürtler, & Schmitz, 2005; Pintrich, Roeser, & De Groot, 1994; Skinner, Furrer, Marchand, & Kindermann, 2008). As a result, negative perceptions of competence lead to decreased engagement in learning activities and underachievement (Bandura, 1993; Fortier, Vallerand, & Guay, 1995; Pajares & Graham, 1999; Shim, Ryan, & Anderson, 2008; B. J. Zimmerman, 2000). In the long run, children and adolescents with lower self-perceived competence and achievement are more likely to drop out of school (Alexander, Entwisle, & Kabbani, 2001; Caprara et al., 2008; Guay, Larose, & Boivin, 2004; Vallerand, Fortier, & Guay, 1997).

Research Objective and Hypothesis

Although research examining outcomes of depression indicates that young people with mental health issues fare poorly in school (Vander Stoep et al., 2002), little is known about the dynamics connecting adolescent depression to school dropout. Theoretical models suggest that depression can interfere with academic experience and lead to dropout through self-perceptions of academic competence and school performance. In this study, we sought to test the mediation linkages proposed. We hypothesized that the relationship between depression symptoms and dropout would be mediated by student self-perceptions of academic competence and self-reported achievement. Further, we anticipated that self-perceived academic competence would influence dropout partly through its association with self-reported achievement.

Method

Participants and Procedure

This study is based on a high-risk longitudinal sample (2000–2006) of French-speaking adolescents living in Montreal (Quebec, Canada; see Morin, Janosz, & Larivée, 2010, for more details). Participants were recruited from two suburban secondary schools ranked by the Ministry of Education of Quebec (MEQ) in the three lowest deciles of socioeconomic status (SES) according to mother's education and parental employment. Many of these students received special education (40%) and had history of grade retention (26%). All the students in seventh grade were invited to take part in the study after obtaining parental consent. Out of 602 students, 496 consented to participate (82.4%). Three students who had not responded to the depression symptoms inventory were withdrawn from the sample. Our final sample included 493 par-

ticipants (228 girls and 265 boys). The average age of participants was 12.54 years (*SD* = 0.73) at the beginning of the study. Most participants came from Caucasian French-speaking families (82%).

The variables used in this study were drawn from data collected when participants were in the seventh grade. Three waves of data collection took place during that school year, in the fall, winter, and spring. Self-reported questionnaires were administered in class with the help of trained research assistants. Some additional information was obtained through the MEQ. We followed this cohort for 6 years to assess student educational enrollment.

Measures

Control variables (parental education, gender, and grade retention) were measured at the beginning of the school year. The predictor variable (depression) was measured in the fall and in spring, and the mediator variables (self-reported academic competence and achievement) were measured at each wave of data collection. For every variable, we calculated the average of the scores gathered at each wave to obtain global measures of participant academic adjustment and well-being in the seventh grade. The outcome (dropout status) was measured up to 6 years later. See Appendix for details of measures.

Outcome variable: School dropout. Secondary education in the province of Quebec spans from the seventh through the 11th grade and is thereby normally completed within 5 years. To determine student dropout status, we followed participants until 1 year beyond expected graduation. This information was obtained through the MEQ’s educational database. The MEQ keeps records of all the students enrolled in public and private schools across the province of Quebec, including those who transferred to a different school, vocational program, or adult education. Although this monitoring system results in some sample bias by labeling students who move out of Quebec as not enrolled, the extent of the bias is limited by low rates of interprovincial mobility across Canada (Bernard, Finnie, & St.-Jean, 2008). We considered students who were continuously enrolled in school or had completed their education and obtained a high school diploma to be nondropouts. Conversely, students who were not enrolled in school a particular year and who had not obtained a diploma were considered dropouts. Overall, 166 students dropped out of school in this study. Among them, 29 dropped out in the 10th grade, 39 dropped out in the 11th grade, and 98 left school without qualification 1 year beyond expected graduation. Although more than a third of dropouts (37%) tried to reenroll during the study, none of them had obtained a diploma by the final follow-up. The dependent variable was coded 0 (nondropout) or 1 (dropout). As shown in Table 1, 33.7% of participants dropped out of school in our sample, and

there were more dropout boys than girls, although this difference reached only marginal statistical significance, $\chi^2 = 3.48(1)$, $p = .07$.

Control variables. To eliminate the effect of potential confounding factors, we controlled for gender, parental education, and grade retention. Participant gender was coded 0 (female) or 1 (male). To assess parental education, we calculated the mean number of years for mother’s education and father’s education reported by participants. Student history of grade retention was measured with the number of retained years in elementary school according to MEQ’s database.

Predictor variable: Depression. Self-reported symptoms of depression were measured with the French version of the Inventory to Diagnose Depression (IDD; M. Zimmerman & Coryell, 1987b), translated by Pariente, Smith, and Guelfi (1989). The 22-item scale can be used to assess depression according to *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) criteria, or to evaluate symptoms of depression on a continuous scale. In this study, we used the continuous scale. Each item is graduated in five propositions, rated from 0 to 4, which illustrate increasing severity of symptoms.

This measure of depression presents remarkably stable psychometric qualities across cultures and populations with adolescents and adults (Ackerson, Weigman Dick, Manson, & Baron, 1990; Ruggero, Johnson, & Cuellar, 2004; Sakado, Sato, Uehara, Sato, & Kameda, 1996). Previous studies have shown that the IDD displays good internal consistency, with a Cronbach’s alpha coefficient of .92 and Spearman–Brown coefficient of .90 (M. Zimmerman & Coryell, 1987a). Similarly, we obtained alpha coefficients ranging from .90 to .93 with our sample. The IDD is also correlated to other known measures of depression such as the Beck Depression Inventory (Beck, 1978), with coefficients ranging from .88 to .96 (Haaga, McDermut, & Ahrens, 1993; Rogers, Adler, Bungay, & Wilson, 2005).

Mediator variables. Student self-perceived academic competence was measured with four items on a Likert-type scale adapted in French from Skinner’s (1995) questionnaire. The items reflected student sense of mastery (“Even if I want to, I can’t succeed in school”; “No matter what I do, I don’t get good grades”) and control in the academic domain (“I get good grades when I want to”; “I can do well in school if I want to”). They were ranked on a scale from 1 (*completely disagree*) to 4 (*completely agree*). Self-reported academic achievement was the mean of student-reported school performance in two basic subjects (language arts and mathematics). Self-reported grades in specific subjects, like mathematics and language arts, are considered to reflect actual grades with reasonable accuracy (Kuncel, Credé, & Thomas, 2005).

Table 1
Distribution of Dropouts and Nondropouts According to Gender

Dropout	Girl		Boy		Total	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Nondropout	161	70.6	166	62.6	327	66.3
Dropout	67	29.4	99	37.4	166	33.7
Total	228	100	265	100	493	100

Analytical Strategy

We tested mediation hypotheses following Baron and Kenny’s (1986) recommendations. According to Baron and Kenny, four conditions are required to establish mediation. There must be a significant relation (a) between the predictor (depression) and the mediator variable (academic competence), (b) between the predictor and the outcome variable (school dropout), and (c) between the mediator and the outcome variable, and (d) the impact of the

predictor on the outcome variable must be altered when controlling for the effect of the mediator.

To support these conditions, first, we tested for correlations to determine the degree of association among variables. Second, we performed a series of simple logistic regressions testing the effect of the predictor and mediator variables on school dropout. Measures were standardized to facilitate the interpretation of coefficients and odds ratios across variables. Odds can thus be interpreted as the expected change in outcome when a variable varies by ± 1 standard deviation and that all other variables are at their sample means.

Third, we tested three models using structural equation modeling (SEM) with weighted least-squares estimation on Mplus 3.13 software (Muthén & Muthén, 2005). The use of SEM allows testing more complex models that include the simultaneous modeling of sequential linkages among variables, testing the strength of direct and indirect effects, and modeling constructs into latent variables. The hypothesized mediation model with paths going from the control variables toward all the predictors and the outcome was initially estimated. Then an alternative model was tested where the path between depression symptoms and academic competence was removed and the direct effect of depression symptoms on school dropout was estimated. Finally, the model with the best fit was estimated as a reduced model with all nonsignificant paths removed. Self-perceived academic competence was modeled as a latent variable with two dimensions measuring student mastery and student control. All other constructs in the model were observed variables. Assessing model fit relies on several indices and commonly used criteria (Hu & Bentler, 1999). Besides the chi-square statistic designed to test how well the model fits the data, we reported the comparative fit index (CFI; Bentler, 1990) and the Tucker–Lewis index (TLI; Tucker & Lewis, 1973) that compare model fit to a more restricted baseline model and require values larger than .95 for good fit. The root-mean-square error of approximation (RMSEA; Steiger, 1990) examining population discrepancy with a recommended cutoff value of .06 was also listed. Standardized coefficients for direct and indirect effects (Bollen, 1989) are reported for the final model.

Results

Correlations Among Study Variables

Descriptive statistics and bivariate correlations among variables are shown in Table 2. Results show that depression scores were

negatively correlated with self-perceived academic competence but not with self-reported academic achievement. As expected, self-reported academic competence and achievement were positively associated.

Prediction of School Dropout

Table 3 presents regression coefficients (*B*) for simple logistic regression analysis showing the degree of association between predictors and the outcome, standard errors, the Wald test and its significance value, and odds ratios (OR). Findings showed that all the variables under study were significantly related to dropping out of school, with the exception of gender, which was marginally significant. The association between depression (*OR* = 1.23) and dropping out of school indicated that adolescents with higher symptoms of depression were 23% more likely to become school dropouts.

Testing the Hypothesized, Alternative, and Reduced Models With SEM

Results of SEM fit indices are presented in Table 4. The chi-square statistic for the hypothesized mediation model was nonsignificant, indicating that the model fitted the data well, $\chi^2 = 6.72(7)$, $p = .46$. All other fit indices reached the expected criteria values: CFI = 1.00, TLI = 1.00, RMSEA = .00. In comparison, the alternative model demonstrated an overall poor fit, with $\chi^2 = 35.67(6)$, $p = .00$, and fit indices below expected values (CFI = .90, TLI = .68, RMSEA = .10), indicating that removing the path between depression and self-perceived academic competence oversimplified the model. The hypothesized mediation model was thereby selected.

To improve model parsimony, some modifications to the hypothesized model were judged necessary in the light of empirical results. A reduced model removing nonsignificant paths from gender to self-perceived competence, and from grade retention to self-reported achievement, was thus estimated. For theoretical considerations, we kept the statistically marginal significant path going from gender to dropping out (standardized coefficient = .08, $p < .15$). All other paths were significant at $p < .05$. The fit values for the reduced model demonstrated adequate fit to the data, $\chi^2 = 6.66(7)$, $p = .47$. CFI = .99 and TLI = .99 showed that the estimated model fit was very good in comparison to an independent model, whereas RMSEA = .00 indicated that the error of approximation in fitting the model to the population covariance

Table 2
Correlations Among Control Variables, Predictor, and Mediators

Variables	1	2	3	4	5	<i>M</i>	<i>SD</i>	Skewness ^a	Kurtosis ^b	Min	Max
1. Parental education	—					2.16	0.97	0.63	−0.58	1	4
2. Grade retention	−.11*	—				0.26	0.44	1.12	1.25	0	1
3. Depression	−.06	.12**	—			14.02	11.03	−0.74	1.80	0	59.50
4. Academic competence	.11*	−.20***	−.22***	—		3.27	0.52	−0.45	−0.52	1.67	4.00
5. Achievement	.19***	−.18***	−.07	.41***	—	72.66	8.85	0.15	−0.59	52.00	95.00

Note. $n = 493$. Min = minimum; max = maximum.

^a $SE = 0.11$. ^b $SE = 0.22$.

* $p < .05$. ** $p < .01$. *** $p < .000$.

Table 3
Prediction of School Dropout With Simple Logistic Regression

Variable	B	SE	Wald	OR
Parental education	−0.33	0.10	10.97**	0.72
Gender	0.36	0.19	3.48 [†]	1.43
Grade retention	1.45	0.22	44.36***	4.25
Depression	0.21	0.09	5.03*	1.23
Academic competence	−0.45	0.10	21.24***	0.64
Achievement	−0.57	0.11	28.93***	0.57

Note. OR = odds ratio.
[†] $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .000$.

matrix was small and that the model fit was good (Browne & Cudeck, 1993).

Examining direct effects. The reduced model showed that adolescents exhibiting symptoms of depression reported feeling less competent and in control in the academic domain (standardized coefficient = $-.23$). Self-perceptions of academic competence were in turn positively related to self-reported achievement (standardized coefficient = $.51$) and negatively to dropping out of school (standardized coefficient = $-.20$). Higher self-reported achievement predicted lower risk of dropping out (standardized coefficient = $-.15$). Girls were more likely to feel depressed (standardized coefficient = $-.32$) and to have better grades than boys (standardized coefficient = $-.09$). Grade retention was negatively associated with most predictors: being retained in elementary school led to feeling more depressed (standardized coefficient = $.14$) and less competent (standardized coefficient = $-.22$) at the beginning of secondary school. It also significantly increased the risk of dropping out (standardized coefficient = $.26$).

Testing indirect effects. We hypothesized that the relationship between depression and dropout would be mediated by self-reported academic competence. This hypothesis was confirmed. The results showed an indirect relationship between depression and dropout through self-perceived academic competence (standardized indirect coefficient = $.05$), indicating that the probability of leaving school before completion was 5% higher each time that depression symptoms increased by 1 standard deviation. Furthermore, the relationship between self-perceived academic competence and dropout was partially mediated by self-reported achievement, with an indirect effect of self-perceptions of competence and dropout through self-reported achievement (standardized indirect coefficient = $-.09$). These findings provided additional evidence of the mediating processes that explain the linkage between depression in adolescents and school leaving.

Table 4
Goodness-of-Fit Indices for Structural Equation Models

Model	χ^2	df	p	CFI	TLI	RMSEA	Free parameters
Hypothesized	6.72	7	.458	1.000	1.000	.000	27
Alternative	35.67	6	.000	.898	.677	.100	27
Reduced	9.60	9	.384	.998	.996	.012	24

Note. Good fit is indicated by nonsignificant chi-square, comparative fit index (CFI) $> .95$, Tucker–Lewis index (TLI) $> .95$, root-mean-square error of approximation (RMSEA) $< .10$.

Discussion

Research on adolescent well-being has suggested that students with depression have an increased probability of facing academic failure (Asarnow et al., 2005), yet few studies have looked at the implications of adolescent depression in the process of school dropout prospectively. This study sought to examine mediating processes implicated in the relationship between depression symptoms and academic adjustment at the entry of secondary school and dropping out in later adolescence. Our findings revealed that self-reported depression symptoms in seventh grade predicted school dropout 1 year beyond expected graduation. The relationship between depression and dropping out of school was, however, mediated by self-perceptions of academic competence. This indicates that depression symptoms at the beginning of secondary school are related to higher dropout mainly by being associated with pessimistic views about the likelihood to reach desired school outcomes; student negative self-beliefs are in turn related to lower self-reported academic performance and predict a higher risk of dropping out. These findings emphasize that the connection between early depression and leaving school without qualifications is mostly indirect, as it is accounted for by achievement-related self-perceptions.

Linking Depression Symptoms to School Dropout

Theoretical models seeking to integrate mental health issues and school experience suggest that depression can lead to academic difficulties through self-regulation processes (Roeser & Eccles, 2000; Rudolph, 2004). Our findings support these models. Students who experience depression are also characterized by negative patterns of self-beliefs and have a higher risk of dropping out partly because they tend to doubt their ability to do well in school. They may also be inclined to ruminate on their negative experiences, feeling anxious and guilty about school performance or blaming themselves for failing in school, which could lead to feelings of helplessness (Eccles, Roeser, Vida, Fredricks, & Wigfield, 2006; Nolen-Hoeksema et al., 1992). The ensuing consequences on school perseverance can be considerable. Individuals who have developed more problematic motivational profiles are more likely to experience decreased engagement and performance (Skinner, Zimmer-Gembeck, & Connell, 1998). Self-perceptions of competence may thus be determining in student success, as they are directly associated with elements of academic engagement and performance that may result in emotional and behavioral withdrawal from school and eventually dropout. Including a broader range of motivational variables (task value, etc.) in future studies

would allow elaborating on the processes linking mental health issues and school success.

The connection between depression and negative self-perceptions did not extend to self-reported academic achievement in this study. Although some studies have reported small but statistically significant associations between depression and academic achievement (Hishinuma et al., 2001; Roeser et al., 1998; Wiest et al., 1998), others have found no relationship between these variables (Fleming et al., 2005; Hamilton et al., 1997; Nurmi et al., 1995). Our findings are consistent with the latter research. In addition to lower perceived competence, children and adolescents with withdrawn behavior or depression also tend to have more helpless school behavior (Nolen-Hoeksema et al., 1992; Roeser et al., 2002). These shortcomings may translate into lower performance mainly because they are related to poor academic self-regulation rather than negative affect per se (Roeser et al., 1998). In time, the association between depression and school performance could become more pronounced as adolescents who are persistently depressed undergo a decline in achievement (Marcotte, Lévesque, & Fortin, 2006). This would suggest a possible long-term effect for those who continuously experience depression. Future studies should investigate whether persistence or fluctuations of depressive mood during adolescence would affect differently school adjustment.

Implications for School Dropout Research

This study underscores the need to include mental health issues in explanatory models of school dropout. Theoretical perspectives on school dropout generally encompass academic, social, and emotional factors, at least to some degree, but for the most part these factors are limited to externalizing problems and deviance (Battin-Pearson et al., 2000; Janosz, Le Blanc, Boulerice, & Tremblay, 2000; Newcomb et al., 2002; Rumberger, 2011). The few studies examining depression in middle adolescence (14–16 years) and later education were unable to demonstrate a relation after controlling for risk factors of dropout (Fergusson & Woodward, 2002; Miech et al., 1999). Fergusson and Woodward (2002) concluded that the association between depression and education was explained by exposure to adversity often co-occurring with depression. The current study, however, indicates that depression in adolescence is one of the links that connect indirectly to lower education. It illustrates the critical need to test for potential mediators between psychological variables and academic failure in order to shed light on the mechanisms involved. Neglecting to examine those indirect links could lead to the underestimation of significant relationships implicated in the process of school dropout.

This study has implications for mental health prevention and school dropout prevention. Mental health prevention should consider negative academic self-perceptions in early adolescence as indicative of potentially coexisting depression symptoms. Students with pessimistic academic self-views ought to be screened for depression and offered appropriate care when indicated. Additionally, interventions that target student mental health and negative self-perceptions are likely to improve dropout prevention. Early interventions that aim at enhancing student mental health and sense of mastery could be instrumental in preventing premature school exit, as they are likely to increase academic engagement

(Appleton, Christenson, Kim, & Reschly, 2006; Christenson & Thurlow, 2004).

Limitations

Since this sample was made up mainly of high-risk French-speaking students from low-SES schools in Montreal, caution should be applied when generalizing the results to other groups. Replication with a normative sample would help to extend findings to the general population. Another limitation of the study is that it relied on self-reported measures of depression, academic competence, and achievement. Our results are thus based on individual perceptions of well-being and academic adjustment. Because large-scale standardized testing is not implemented in Quebec schools, we used student reports of achievement in specific subjects (language arts and mathematics), which tend to be more reliable than self-reported GPA (Kuncel et al., 2005), but ideally future studies should examine the effect of actual GPA. Third, the mediation analyses were based on seventh-grade measures of adolescent experience (depression, self-reported academic competence, and achievement), therefore restricting our observations to the beginning of secondary school. Although we accounted for some aspects of earlier school experience, such as history of grade retention, other studies should consider more diverse measures of academic experience and emotional well-being in primary school. Despite these caveats, this study encompasses a number of strengths worth mentioning. It was based on a 6-year prospective design that allowed us to draw conclusions about the long-term consequences of adolescent depression symptoms and academic experience on subsequent educational attainment. Moreover, whereas previous research on depression and dropout relied on observations in middle adolescence (14–16 years; Fergusson & Woodward, 2002; Miech et al., 1999), we investigated outcomes of depression based on information gathered at 12 years of age, thus enabling our understanding of the consequences of early adolescent experience on schooling.

Conclusion

The implications of this study shed light on early mechanisms involved in the process of dropping out. School dropout is a cumulative process influenced by multiple risk factors (Rumberger, 2011). Whereas among individual factors, achievement-related variables have received much attention, emotional factors have been studied unevenly. This study brings attention to adolescent depression symptoms, a risk factor largely overlooked by research in the past, and shows its ties to school dropout mainly through the association with achievement-related beliefs. In a context of school reform in the United States and in Canada, this informs policy and practice about the necessity to improve school-based mental health promotion and services to enhance adolescent well-being and academic success.

References

- Abramson, L. Y., Metalsky, G. I., & Alloy, L. B. (1989). Hopelessness depression: A theory-based subtype of depression. *Psychological Review*, 96, 358–372. doi:10.1037/0033-295X.96.2.358
- Ackerson, L. M., Weigman Dick, R., Manson, S. M., & Baron, A. E. (1990). Properties of the Inventory to Diagnose Depression in American

- Indian adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 29, 601–607. doi:10.1097/00004583-199007000-00014
- Alexander, K. L., Entwisle, D. R., & Kabbani, N. S. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record*, 103, 760–822. doi:10.1111/0161-4681.00134
- Aluja, A., & Blanch, A. (2002). The Children Depression Inventory as predictor of social and scholastic competence. *European Journal of Psychological Assessment*, 18, 259–274. doi:10.1027//1015-5759.18.3.259
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychosocial engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, 44, 427–445. doi:10.1016/j.jsp.2006.04.002
- Asarnow, J. R., Jaycox, L. H., Duan, N., LaBorde, A. P., Rea, M. M., Tang, L., . . . Wells, K. B. (2005). Depression and role impairment among adolescents in primary care clinics. *Journal of Adolescent Health*, 37, 477–483. doi:10.1016/j.jadohealth.2004.11.123
- Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, 28, 117–148. doi:10.1207/s15326985ep2802_3
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182. doi:10.1037/0022-3514.51.6.1173
- Battin-Pearson, S., Newcomb, M. D., Abbott, R. D., Hill, K. G., Catalano, R. F., & Hawkins, J. D. (2000). Predictors of early high school dropout: A test of five theories. *Journal of Educational Psychology*, 92, 568–582. doi:10.1037/0022-0663.92.3.568
- Beck, A. T. (1978). *Depression Inventory*. Philadelphia, PA: Center for Cognitive Therapy.
- Bentler, P. M. (1990). Comparative fit indexes in structural equation models. *Psychological Bulletin*, 107, 238–246. doi:10.1037/0033-2909.107.2.238
- Bernard, A., Finnie, R., & St.-Jean, B. (2008). Interprovincial mobility and earnings. *Perspectives on Labour and Income*, 9(10), 15–25.
- Birmaher, B., Bridge, J. A., Williamson, D. E., Brent, D. A., Dahl, R. E., Axelson, D. A., . . . Ryan, N. D. (2004). Psychosocial functioning in youths at high risk to develop major depressive disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43, 839–846. doi:10.1097/01.chi.0000128787.88201.1b
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bouffard-Bouchard, T., Parent, S., & Larivée, S. (1991). Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students. *International Journal of Behavioral Development*, 14, 153–164.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury, CA: Sage.
- Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Barbaranelli, C., & Bandura, A. (2008). Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology*, 100, 525–534. doi:10.1037/0022-0663.100.3.525
- Christenson, S. L., & Thurlow, M. L. (2004). School dropouts: Prevention considerations, interventions, and challenges. *Current Directions in Psychological Science*, 13, 36–39. doi:10.1111/j.0963-7214.2004.01301010.x
- Cole, D. A., Martin, J. M., Peeke, L. A., Seroczynski, A. D., & Fier, J. (1999). Children's over- and underestimation of academic competence: A longitudinal study of gender differences, depression, and anxiety. *Child Development*, 70, 459–473. doi:10.1111/1467-8624.00033
- Cole, D. A., Martin, J. M., & Powers, B. (1997). A competency-based model of child depression: A longitudinal study of peer, parent, teacher, and self-evaluations. *Journal of Child Psychology and Psychiatry*, 38, 505–514. doi:10.1111/j.1469-7610.1997.tb01537.x
- Eccles, J. S., Roeser, R., Vida, M., Fredricks, J., & Wigfield, A. (2006). Motivational and achievement pathways through middle childhood. In L. Balter & C. S. Tamis-LeMonda (Eds.), *Child psychology: A handbook of contemporary issues* (pp. 325–355). New York, NY: Psychology Press.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132. doi:10.1146/annurev.psych.53.100901.135153
- Fergusson, D. M., & Woodward, L. J. (2002). Mental health, educational, and social role outcomes of adolescents with depression. *Archives of General Psychiatry*, 59, 225–231. doi:10.1001/archpsyc.59.3.225
- Fleming, C. B., Haggerty, K. P., Catalano, R. F., Harachi, T. W., Mazza, J. J., & Gruman, D. H. (2005). Do social and behavioral characteristics targeted by preventive interventions predict standardized test scores and grades? *Journal of School Health*, 75, 342–349.
- Fortier, M. S., Vallerand, R. J., & Guay, F. (1995). Academic motivation and school performance: Toward a structural model. *Contemporary Educational Psychology*, 20, 257–274. doi:10.1006/ceps.1995.1017
- Guay, F., Larose, S., & Boivin, M. (2004). Academic self-concept and educational attainment level: A ten-year longitudinal study. *Self and Identity*, 3, 53–68. doi:10.1080/13576500342000040
- Haaga, D. A., McDermut, W., & Ahrens, A. H. (1993). Discriminant validity of the Inventory to Diagnose Depression. *Journal of Personality Assessment*, 60, 285–289. doi:10.1207/s15327752jpa6002_6
- Hamilton, E. B., Asarnow, J. R., & Tompson, M. C. (1997). Social, academic, and behavioral competence of depressed children: Relationship to diagnostic status and family interaction style. *Journal of Youth and Adolescence*, 26, 77–87. doi:10.1023/A:1024592213017
- Hankin, B. L., Abramson, L. Y., & Siler, M. (2001). A prospective test of the hopelessness theory of depression in adolescence. *Cognitive Therapy and Research*, 25, 607–632. doi:10.1023/A:1005561616506
- Hishinuma, E. S., Foster, J. E., Miyamoto, R. H., Nishimura, S. T., Andrade, N. N., Nahulu, L. B., . . . Carlton, B. S. (2001). Association between measures of academic performance and psychosocial adjustment for Asian/Pacific-Islander adolescents. *School Psychology International*, 22, 303–319. doi:10.1177/0143034301223007
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, 73, 509–527. doi:10.1111/1467-8624.00421
- Janosz, M., Le Blanc, M., Boulerice, B., & Tremblay, R. E. (2000). Predicting different types of school dropouts: A typological approach with two longitudinal samples. *Journal of Educational Psychology*, 92, 171–190. doi:10.1037/0022-0663.92.1.171
- Kovacs, M., & Goldstone, D. (1991). Cognitive and social cognitive development of depressed children and adolescents. *Journal for the American Academy of Child & Adolescent Psychiatry*, 30, 388–392. doi:10.1097/00004583-199105000-00006
- Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, 75, 63–82. doi:10.3102/00346543075001063
- Malmberg, L.-E., Wanner, B., & Little, T. D. (2008). Age and school-type differences in children's beliefs about school performance. *International*

- Journal of Behavioral Development*, 32, 531–541. doi:10.1177/0165025408095558
- Marcotte, D., Lévesque, N., & Fortin, L. (2006). Variations of cognitive distortions and school performance in depressed and non-depressed high school adolescents: A two-year longitudinal study. *Cognitive Therapy and Research*, 30, 211–225. doi:10.1007/s10608-006-9020-2
- Miech, R. A., Caspi, A., Moffitt, T. E., Wright, B. R. E., & Silva, P. A. (1999). Low socioeconomic status and mental disorders: A longitudinal study of selection and causation during young adulthood. *American Journal of Sociology*, 104, 1096–1131. doi:10.1086/210137
- Morin, A. J. S., Janosz, M., & Larivée, S. (2010). The Montreal Adolescent Depression Development Project (MADDP): School life and depression following high school transition. *Psychiatry Research Journal*, 1, 183–232.
- Muris, P., Schouten, E., Meesters, C., & Gijsbers, H. (2003). Contingency-competence-control-related beliefs and symptoms of anxiety and depression in a young adolescent sample. *Child Psychiatry and Human Development*, 33, 325–339. doi:10.1023/A:1023040430308
- Muthén, L. K., & Muthén, B. O. (2005). *Mplus* (Version 3.13) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Newcomb, M. D., Abbott, R. D., Catalano, R. F., Hawkins, J. D., Battin-Pearson, S., & Hill, K. (2002). Mediation and deviance theories of late high school failure: Process roles of structural strains, academic competence, and general versus specific problem behavior. *Journal of Counseling Psychology*, 49, 172–186. doi:10.1037/0022-0167.49.2.172
- Nolen-Hoeksema, S., Girgus, J. S., & Seligman, M. E. P. (1992). Predictors and consequences of childhood depressive symptoms: A 5-year longitudinal study. *Journal of Abnormal Psychology*, 101, 405–422. doi:10.1037/0021-843X.101.3.405
- Nurmi, J.-E., Onatsu, T., & Haavisto, T. (1995). Underachievers' cognitive and behavioral strategies—Self-handicapping at school. *Contemporary Educational Psychology*, 20, 188–200. doi:10.1006/ceps.1995.1012
- Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school. *Contemporary Educational Psychology*, 24, 124–139. doi:10.1006/ceps.1998.0991
- Pariente, P., Smith, M., & Guelfi, J.-D. (1989). Un questionnaire pour le diagnostic d'épisode dépressif majeur: L'inventaire pour le diagnostic de la dépression (IDD). Présentation de la version française [A questionnaire to diagnose major depressive episode: The French Inventory to Diagnose Depression (IDD)]. *Psychiatrie et Psychobiologie*, 4, 375–385.
- Perels, F., Gürtler, T., & Schmitz, B. (2005). Training of self-regulatory and problem-solving competence. *Learning and Instruction*, 15, 123–139. doi:10.1016/j.learninstruc.2005.04.010
- Pintrich, P. R., Roeser, R. W., & De Groot, E. A. (1994). Classroom and individual differences in early adolescents' motivation and self-regulated learning. *Journal of Early Adolescence*, 14, 139–161. doi:10.1177/027243169401400204
- Puig-Antich, J., Kaufman, J., Ryan, N. D., Williamson, D. E., Dahl, R. E., Lukens, E., . . . Nelson, B. (1993). The psychosocial functioning and family environment of depressed adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 32, 244–253. doi:10.1097/00004583-199303000-00003
- Quiroga, C. V., Janosz, M., Lyons, J. S., & Morin, A. J. S. (2012). Grade retention and seventh-grade depression symptoms in the course of school dropout among high-risk adolescents. *Psychology*, 3, 749–755. doi:10.4236/psych.2012.329113
- Reinherz, H. Z., Frost, A. K., & Pakiz, B. (1991). Changing faces: Correlates of depressive symptoms in late adolescence. *Family and Community Health*, 14, 52–63.
- Roeser, R. W., & Eccles, J. S. (2000). Schooling and mental health. In A. J. Sameroff, M. Lewis, & S. M. Miller (Eds.), *Handbook of developmental psychopathology* (2nd ed., pp. 135–156). New York, NY: Springer. doi:10.1007/978-1-4615-4163-9_8
- Roeser, R. W., Eccles, J. S., & Sameroff, A. J. (1998). Academic and emotional functioning in early adolescence: Longitudinal relations, patterns, and prediction by experience in middle school. *Development and Psychopathology*, 10, 321–352. doi:10.1017/S0954579498001631
- Roeser, R. W., Eccles, J. S., & Sameroff, A. J. (2000). School as a context of early adolescents' academic and social-emotional development: A summary of research findings. *Elementary School Journal*, 100, 443–471. doi:10.1086/499650
- Roeser, R. W., Strobel, K. R., & Quihuis, G. (2002). Studying early adolescents' academic motivation, social-emotional functioning and engagement in learning: Variable- and person-centered approaches. *Anxiety, Stress & Coping*, 15, 345–368. doi:10.1080/1061580021000056519
- Rogers, W. H., Adler, D. A., Bungay, K. M., & Wilson, I. B. (2005). Depression screening instruments made good severity measures in a cross-sectional analysis. *Journal of Clinical Epidemiology*, 58, 370–377. doi:10.1016/j.jclinepi.2004.10.010
- Rudolph, K. D. (2004). A self-regulation approach to understanding adolescent depression in the school context. In T. Urdan & F. Pajares (Eds.), *Educating adolescents: Challenges and strategies* (pp. 33–64). Greenwich, CT: Information Age.
- Ruggero, C. J., Johnson, S. L., & Cuellar, A. K. (2004). Spanish-language measures of mania and depression. *Psychological Assessment*, 16, 381–385. doi:10.1037/1040-3590.16.4.381
- Rumberger, R. W. (2011). *Dropping out: Why students drop out of high school and what can be done about it*. Cambridge, MA: Harvard University Press. doi:10.4159/harvard.9780674063167
- Sakado, K., Sato, T., Uehara, T., Sato, S., & Kameda, K. (1996). Discriminant validity of the Inventory to Diagnose Depression, Lifetime version. *Acta Psychiatrica Scandinavica*, 93, 257–260. doi:10.1111/j.1600-0447.1996.tb10644.x
- Shahar, G., Henrich, C. C., Winokur, A., Blatt, S. J., Kuperminc, G. P., & Leadbeater, B. J. (2006). Self-criticism and depressive symptomatology interact to predict middle school academic achievement. *Journal of Clinical Psychology*, 62, 147–155. doi:10.1002/jclp.20210
- Shim, S. S., Ryan, A. M., & Anderson, C. J. (2008). Achievement goals and achievement during early adolescence: Examining time-varying predictor and outcome variables in growth-curve analysis. *Journal of Educational Psychology*, 100, 655–671. doi:10.1037/0022-0663.100.3.655
- Skinner, E. A. (1995). *Perceived control, motivation, and coping*. Thousand Oaks, CA: Sage.
- Skinner, E. A., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and dissatisfaction in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, 100, 765–781. doi:10.1037/a0012840
- Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology*, 82, 22–32. doi:10.1037/0022-0663.82.1.22
- Skinner, E. A., Zimmer-Gembeck, M. J., & Connell, J. P. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development*, 63(2–3, Serial No. 254).
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval approach. *Multivariate Behavioral Research*, 25, 173–180. doi:10.1207/s15327906mbr2502_4
- Thompson, E. A., Moody, K. A., & Eggert, L. L. (1994). Discriminating suicide ideation among high-risk youth. *Journal of School Health*, 64, 361–367. doi:10.1111/j.1746-1561.1994.tb06205.x
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10. doi:10.1007/BF02291170

Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology*, 72, 1161–1176. doi:10.1037/0022-3514.72.5.1161

Vander Stoep, A., Weiss, N. S., McKnight, B., Beresford, S. A. A., & Cohen, P. (2002). Which measure of adolescent psychiatric disorder—diagnosis, number of symptoms, or adaptive functioning—best predicts adverse young adult outcomes? *Journal of Epidemiology & Community Health*, 56, 56–65. doi:10.1136/jech.56.1.56

Wagner, M., Kutash, K., Duchnowski, A. J., Epstein, M. H., & Sumi, C. W. (2005). The children and youth we serve: A national picture of the characteristics of students with emotional disturbances receiving special education. *Journal of Emotional and Behavioral Disorders*, 13, 79–96. doi:10.1177/10634266050130020201

Wiest, D. J., Wong, E. H., & Kreil, D. A. (1998). Predictors of global self-worth and academic performance among regular education, learning disabled, and continuation high school students. *Adolescence*, 33, 601–618.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25, 82–91. doi:10.1006/ceps.1999.1016

Zimmerman, M., & Coryell, W. (1987a). The Inventory to Diagnose Depression, lifetime version. *Acta Psychiatrica Scandinavica*, 75, 495–499. doi:10.1111/j.1600-0447.1987.tb02824.x

Zimmerman, M., & Coryell, W. (1987b). The Inventory to Diagnose Depression (IDD): A self-report scale to diagnose major depressive disorder. *Journal of Consulting and Clinical Psychology*, 55, 55–59. doi:10.1037/0022-006X.55.1.55

Appendix

Description of the Measures Used in the Study

Variable	Wave	Item	Alpha	Item sample	Scale
School dropout				Status in 10th–11th+ grade according to MEQ	0 (nondropout), 1 (dropout)
Gender	1	1		What is your gender?	0 (girl), 1 (boy)
Parental education	1	2		Which level of education has your mother reached?	1 (<i>secondary education not completed</i>) to 4 (<i>university enrollment</i>)
Grade retention	1			History of retention in elementary school (MEQ)	0 (never retained), 1 (retained)
Academic competence	1–3 ^a	4	.73–.81	No matter what I do, I don’t get good grades. ^b	1 (<i>completely disagree</i>) to 4 (<i>completely agree</i>)
Achievement	1–3 ^a	2		What is your average grade in mathematic?	0%–100%
Depression	1 and 3 ^a	22	.90–.93	(a) I was not sleeping less than normal; (b) I occasionally had slight difficulty sleeping; (c) I clearly didn’t sleep as well as usual; (d) I slept about half my normal amount of time; (e) I slept less than two hours per night.	Symptom count: 0–88

Note. MEQ = Ministry of Education of Quebec.
^a Used the mean score for waves. ^b Inversed item was recoded.

Received January 25, 2010
Revision received December 10, 2012
Accepted December 13, 2012 ■

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, 139, 133–151. <http://dx.doi.org/10.1037/a0028566>
- Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied as Tiff, EPS, or PowerPoint. The minimum line weight for line art is 0.5 point for optimal printing. Original color figures can be printed in color at the editor's and publisher's discretion provided the author agrees to pay \$255 for one figure, \$425 for two figures, \$575 for three figures, \$675 for four figures, and \$55 for each additional figure.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult

journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . ." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . ." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers are required to obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including, for example, test materials or portions thereof and photographs of people.

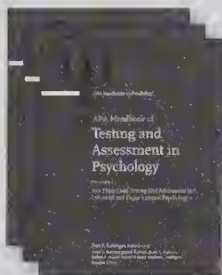
Supplemental materials. APA can now place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). Authors should keep a copy of the manuscript to guard against loss. General correspondence may be addressed to the editorial office at jedgar@memphis.edu.

Preparing files for production. If your manuscript is accepted for publication, please follow the guidelines for file formats and naming provided at www.apa.org/pubs/journals/authors/preparing-efiles.aspx. If your manuscript was mask reviewed, please ensure that the final version for production includes a byline and full author note for typesetting.

NEW RELEASES

from the American Psychological Association



APA Handbook of Testing and Assessment in Psychology

Volume 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology

Volume 2. Testing and Assessment in Clinical and Counseling Psychology

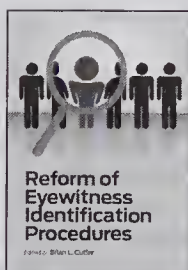
Volume 3. Testing and Assessment in School Psychology and Education

Editor in Chief Kurt F. Geisinger

Series: APA Handbooks in Psychology™

2013. 2,220 pages. 3-Volume Set

List \$695.00 | APA Member/Affiliate: \$395.00
ISBN 978-1-4338-1227-9 | Item # 4311510

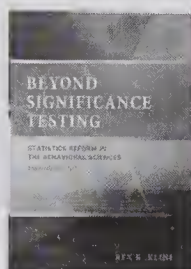


Reform of Eyewitness Identification Procedures

Edited by Brian L. Cutler

2013. 232 pages. Hardcover.

List \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1283-5 | Item # 4318116



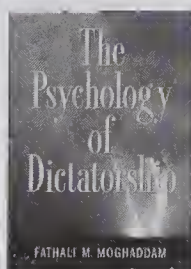
Beyond Significance Testing

Statistics Reform in the Behavioral Sciences
SECOND EDITION

Rex B. Kline

2013. 328 pages. Hardcover.

List \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1278-1 | Item # 4316151

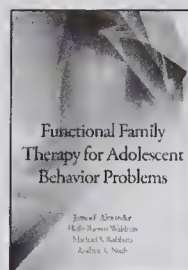


The Psychology of Dictatorship

Fathali M. Moghaddam

2013. 264 pages. Hardcover.

List \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1298-9 | Item # 4316153



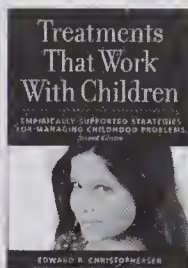
Functional Family Therapy for Adolescent Behavior Problems

An Evidence-Based Approach to Emotional and Behavioral Disorders

James F. Alexander, Holly Barrett Waldron, Michael S. Robbins, and Andrea A. Neeb

2013. 240 pages. Hardcover.

List \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1294-1 | Item # 4317302



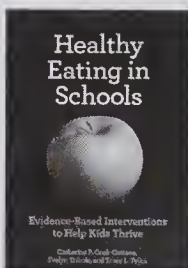
Treatments That Work With Children

Empirically Supported Strategies for Managing Childhood Problems
SECOND EDITION

Edward R. Christophersen and Susan Mortweet VanScoyoc

2013. 328 pages. Hardcover.

List \$49.95 | APA Member/Affiliate: \$39.95
ISBN 978-1-4338-1304-7 | Item # 4317305

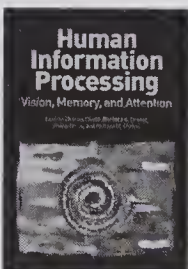


Healthy Eating in Schools

Evidence-Based Interventions to Help Kids Thrive
Catherine P. Cook-Cottone, Evelyn Tribole, and Tracy L. Tylka

2013. 288 pages. Hardcover.

List \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1300-9 | Item # 4317303



Human Information Processing

Vision, Memory, and Attention

Edited by Charles Chubb, Barbara A. Doshier, Zhong-Lin Lu, and Richard M. Shiffrin

2013. 367 pages. Paperback.

List \$29.95 | APA Member/Affiliate: \$24.95
ISBN 978-1-4338-1296-5 | Item # 4313036



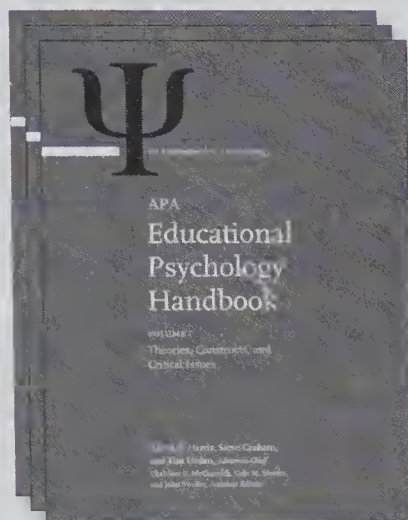
AMERICAN PSYCHOLOGICAL ASSOCIATION

TO ORDER: 800-374-2721 • www.apa.org/pubs/books

AD2244

APA EDUCATIONAL PSYCHOLOGY HANDBOOK

Editors-in-Chief Karen R. Harris, Steve Graham,
and Tim Urdan



The *APA Educational Psychology Handbook* reflects the broad nature of the field today, with state-of-the-science reviews of the diverse critical theories driving research and practice; in-depth investigation of the range of individual differences and cultural/contextual factors that affect student achievement, motivation, and beliefs; and close examination of the research driving current assessment, decision making, teaching skills and content, teacher preparation, and the promotion of learning across the life span and with special populations.

- Volume 1 addresses the definition of educational psychology, some of the most critical theories driving research and practice today, broad areas of research that educational psychology has addressed based on multiple theories and that make an important contribution to the field, and emerging and cutting-edge issues.
- Volume 2 includes 21 chapters that examine a range of individual differences, cultural factors, and contextual factors affecting student achievement, motivation, and beliefs.
- Volume 3 focuses on specific applications of research in educational psychology for assessment and decision making, teaching skills and content, promoting learning, and teacher preparation as well as across the life span and with special populations.

2012. 1,843 pages. 3-Volume Set.

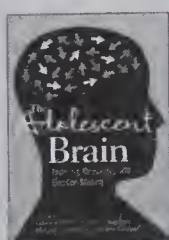
.....
List: \$595.00 | APA Member/Affiliate: \$295.00 | ISBN 978-1-4338-0996-5 | Item # 4311503

CONTENTS

For a detailed Table of Contents, including a list of chapter titles and contributors, please visit the book's online site at www.apa.org/pubs/books/4311503.aspx.

Volume 1: Theories, Constructs, and Critical Issues | I. Conceptualization, Research Design, and Foundational Theories | II. Theory and Research on Critical Topics: What We Know and Why It Matters | III. Emerging Issues and Cutting-Edge Topics |
Volume 2: Individual Differences and Cultural and Contextual Factors | I. Individual Differences | II. Instructional Influences on Motivation, Engagement, Conceptual Change, and Moral Development | III. Cultural and Neighborhood Effects | IV. Relationships | V. Teachers and Classroom Contexts | **Volume 3: Application to Learning and Teaching** | I. Application Across the Life Span | II. Assessment and Decision Making in Education | III. Teaching Core Skills and Content | IV. Instructional Methods | V. Teaching Special Populations

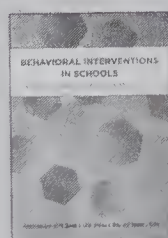
ALSO OF INTEREST



The Adolescent Brain
Learning, Reasoning, and Decision Making
Edited by Valerie

F. Reyna, Sandra B. Chapman,
Michael R. Dougherty,
and Jere Confrey
2012. 440 pages. Hardcover.

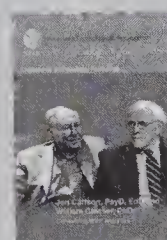
.....
List: \$79.95 | APA Member/Affiliate: \$59.95
ISBN 978-1-4338-1070-1 | Item # 4318098



Behavioral Interventions in Schools
Evidence-Based Positive Strategies
Edited by
Angeleque Akin-

Little, Steven G. Little, Melissa A.
Bray, and Thomas J. Kehle
2009. 350 pages. Hardcover.

.....
List: \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-0460-1 | Item # 4317189
CEP Credit: 8



Consulting With Teachers
with Jon Carlson
and William Glasser
DVD. Over 100 minutes.

.....
List: \$99.95 | APA Member/Affiliate: \$69.95
ISBN 978-1-59147-352-7 | Item # 4310735



AMERICAN PSYCHOLOGICAL ASSOCIATION

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2004

BEST SELLERS

from the American Psychological Association



AMERICAN
PSYCHOLOGICAL
ASSOCIATION

APA Handbook of Testing and Assessment in Psychology

Volume 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology

Volume 2. Testing and Assessment in Clinical and Counseling Psychology

Volume 3. Testing and Assessment in School Psychology and Education

Editor-in-Chief Kurt F. Geisinger

2013. 2,220 pages. Hardcover.

Series: *APA Handbooks in Psychology™*

List: \$695.00

APA Member/Affiliate: \$395.00

ISBN 978-1-4338-1227-9 | Item # 4311510

APA Handbook of Psychology, Religion, and Spirituality

Volume 1. Context, Theory, and Research

Volume 2. An Applied Psychology of Religion and Spirituality

Editor-in-Chief Kenneth I. Pargament

2013. 1,496 pages. Hardcover.

Series: *APA Handbooks in Psychology™*

List: \$495.00

APA Member/Affiliate: \$249.00

ISBN 978-1-4338-1077-0 | Item # 4311506

Your Complete Guide to College Success

How to Study Smart, Achieve Your Goals, and Enjoy Campus Life

Donald J. Foss

2013. 352 pages. Paperback.

List: \$29.95 | APA Member/Affiliate: \$24.95

ISBN 978-1-4338-1296-5 | Item # 4313036

Purpose and Meaning in the Workplace

Edited by Bryan J. Dik, Zinta S. Byrne, and Michael F. Steger

2013. 240 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1314-6 | Item # 4318117

Beyond Significance Testing

Statistics Reform in the Behavioral Sciences

SECOND EDITION

Rex B. Kline

2013. 440 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1278-1 | Item # 4316151

Human Information Processing

Vision, Memory, and Attention

Edited by Charles Chubb,

Barbara A. Doshier, Zhong-Lin Lu,

and Richard M. Shiffrin

2013. 427 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1273-6 | Item # 4318115

The Psychology of Dictatorship

Fathali M. Moghaddam

2013. 264 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1298-9 | Item # 431615

Treatments That Work With Children

Empirically Supported Strategies for Managing Childhood Problems

SECOND EDITION

Edward R. Christophersen and

Susan Mortweet Vanscoyoc

2013. 328 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95

ISBN 978-1-4338-1304-7 | Item # 4317305

Functional Family Therapy for Adolescent Behavior Problems

James F. Alexander, Holly Barrett Waldron,

Michael S. Robbins, and Andrea A. Neeb

2013. 312 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1294-1 | Item # 4317302

Neuropsychological Assessment and Intervention for Youth

An Evidence-Based Approach to Emotional and Behavioral Disorders

Edited by Linda A. Reddy,

Adam S. Weissman, and James B. Hale

2013. 344 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1266-8 | Item # 4316149

Healthy Eating in Schools

Evidence-Based Interventions to Help Kids Thrive

Catherine P. Cook-Cottone, Evelyn Tribole,

and Tracy L. Tylka

2013. 288 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1300-9 | Item # 4317303

Families of Children With Developmental Disabilities

Understanding Stress and Opportunities for Growth

David W. Carroll

2013. 240 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1329-0 | Item # 4316155

That's So Gay!

Microaggressions and the Lesbian, Gay, Bisexual, and Transgender Community

Kevin L. Nadal

2013. 288 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1280-4 | Item # 4316152

Cognition and Brain Development

Converging Evidence From Various Methodologies

Edited by Bhoomika Rastogi Kar

2013. 328 pages. Hardcover.

Series: *Human Brain Development*

List: \$79.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1271-2 | Item # 4318114

Internet Sex Offenders

Michael C. Seto

2013. 320 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1364-1 | Item # 4316156

Principles of Forensic Report Writing

Michael Karson and Lavita Nadkarni

2013. 208 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1306-1 | Item # 4317306

Strategic Decision Making in Cognitive Behavioral Therapy

Amy Wenzel

2013. 320 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$44.95

ISBN 978-1-4338-1319-1 | Item # 4317308

The Suicidal Patient

Clinical and Legal Standards of Care

THIRD EDITION

Bruce Bongar and Glenn R. Sullivan

2013. 416 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1325-2 | Item # 4317307

Attachment-Based Psychotherapy

Helping Patients Develop Adaptive Capacities

Peter C. Costello

2013. 264 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1302-3 | Item # 4317304

Attachment in Group Psychotherapy

Cheri L. Marmarosh, Rayna D. Markin,

and Eric B. Spiegel

2013. 296 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1321-4 | Item # 4317309

Building Psychological Resilience in Military Personnel

Theory and Practice

Edited by Robert Sinclair and Thomas Britt

2013. 256 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1331-3 | Item # 4317311

Reform of Eyewitness Identification Procedures

Edited by Brian L. Cutler

2013. 232 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1283-5 | Item # 4318116

The Best Within Us

Positive Psychology Perspectives on Eudaimonia

Edited by Alan S. Waterman

2013. 304 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$59.95

ISBN 978-1-4338-1261-3 | Item # 4318113

Internationalizing Multiculturalism

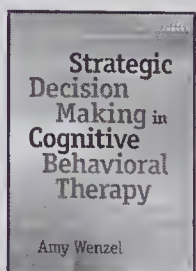
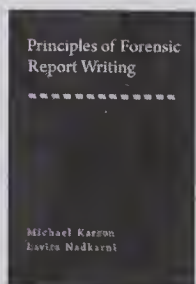
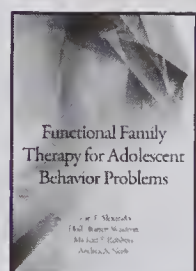
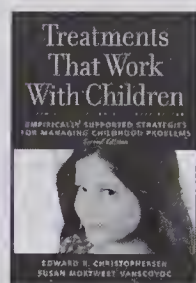
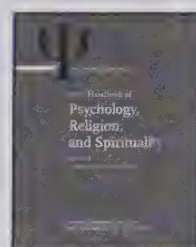
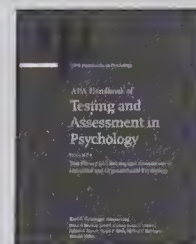
Expanding Professional Competencies in a Globalized World

Edited by Rodney L. Lowman

2013. 338 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

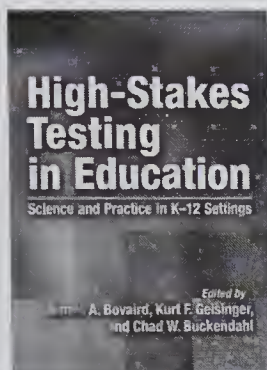
ISBN 978-1-4338-1259-0 | Item # 4317299



HIGH-STAKES TESTING IN EDUCATION

Science and Practice in K-12 Settings

Edited by James A. Bovaird, Kurt F. Geisinger,
and Chad W. Buckendahl



Educational assessment and, more broadly, educational research in the United States have entered into an era characterized by a dramatic increase in the prevalence and importance of test score use in accountability systems.

This volume covers a selection of contemporary issues about testing science and practice that impact the nation's public education system, including local and state assessment development, assessing special populations, charter schools, and the role of college placement and entrance examinations. Also featured is a section focusing on validation practices, defining, and interpreting resulting test scores. Specific topics include the role of examinee motivation, obtaining and making decisions based on validity evidence, evidence of consequences, and considering contextual sampling effects when evaluating validity evidence.

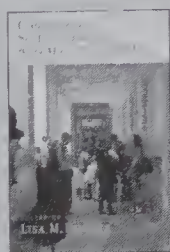
This text is for measurement practitioners, scholars, and advanced graduate students involved in researching and implementing practice and policies for high stakes testing. It will serve as a valuable reference for practitioners and an excellent resource for graduate level seminars in high stakes testing. 2011. 264 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-0973-6 | Item # 4318089

Contents

Introduction, James A. Bovaird, Kurt F. Geisinger, and Chad W. Buckendahl | 1. Current State of High-Stakes Testing in Education, Barbara S. Plake | I. **Current Issues in Kindergarten Through Grade 12 Assessment** | 2. Local Development of a High-Stakes Assessment Program: Lessons Learned and Research Results Gained, John Crawford and Patricia Crum | 3. Setting Performance Standards on Alternate Assessments for Students With Disabilities, Abdullah A. Ferdous, Sue Bechard, and Chad W. Buckendahl | 4. Assessing English Language Skills in Education: Implications for Students With Limited English Proficiency, Anja Römhild and James A. Bovaird | 5. Student Achievement and Adequate Yearly Progress Within the Indiana Charter School System, W. Holmes Finch, Brian F. French, and Mary Baker | 6. Revising a Large-Scale College Placement Examination Program: Innovation Within Constraints, Kristen Huff and Gerald J. Melican | 7. A Role for College Admissions Tests in State Assessment Programs, Sherri Miller and Jay Happel | II. **Validity Considerations: Test Use and Consequences of Test Scores** | 8. Finding Validity Evidence: An Analysis Using the Mental Measurements Yearbook, Gregory J. Cizek, Heather K. Koons, and Sharyn L. Rosenberg | 9. A Model of Examinee Test-Taking Effort, Steven L. Wise and Lisa F. Smith | 10. Validity Arguments for High-Stakes Testing and Accountability Systems, Deborah L. Bandalos, Amanda E. Ferster, Susan L. Davis, and Karen M. Samuelsen | 11. Testing and Measurement From a Multilevel View: Psychometrics and Validation, Bruno D. Zumbo and Barry Forer | 12. A High-Stakes Use of Intelligence Testing: A Forensic Case Study, Chad W. Buckendahl and Brett P. Foley | 13. High-Stakes Education Research: Enhanced Scrutiny of the Education System and Its Consequences, James A. Bovaird and Natalie A. Kozioł | III. **Looking Ahead** | 14. The Future of High-Stakes Testing in Education, Kurt F. Geisinger

ALSO OF INTEREST



Conducting Science-Based Psychology Research in Schools

Edited by
Lisa M. Dinella

2009. 225 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95
ISBN 978-1-4338-0468-7 | Item # 4317197

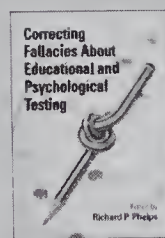


Methodologies for Conducting Research on Giftedness

Edited by
Bruce Thompson
and Rena F. Subotnik

2010. 266 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-0714-5 | Item # 4318068



Correcting Fallacies About Educational and Psychological Testing

Edited by
Richard P. Phelps

2009. 287 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95
ISBN 978-1-4338-0392-5 | Item # 4318046



AMERICAN PSYCHOLOGICAL ASSOCIATION

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

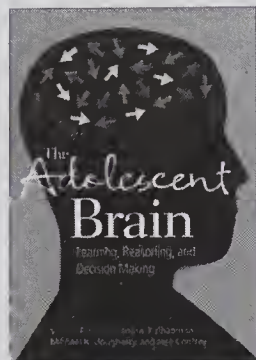
In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD1035

THE ADOLESCENT BRAIN

Learning, Reasoning, and Decision Making

Edited by Valerie F. Reyna, Sandra B. Chapman,
Michael R. Dougherty, and Jere Confrey



The period from adolescence through young adulthood is one of great promise and vulnerability. As teenagers approach maturity, they must develop and apply the skills and habits necessary to navigate adulthood and compete in an ever more technological and globalized world. But as parents and researchers have long known, there is a crucial dichotomy between adolescents' cognitive competence and their frequent inability to utilize that competence in everyday decision-making.

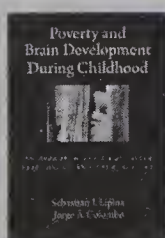
This volume brings together an interdisciplinary group of leading scientists to examine how the adolescent brain develops, and how this development impacts various aspects of reasoning and decision-making, from the use and function of memory and representation, to judgment, mathematical problem-solving, and the construction of meaning. 2012. 440 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$59.95 | ISBN 978-1-4338-1070-1 | Item # 4318098

CONTENTS

Introduction, Valerie F. Reyna, Sandra B. Chapman, Michael R. Dougherty, and Jere Confrey | I. **Foundations** | 1. Anatomic Magnetic Resonance Imaging of the Developing Child and Adolescent Brain, Jay N. Giedd, Michael Stockman, Catherine Weddle, Maria Liverpool, Gregory L. Wallace, Nancy R. Lee, Francois Lalonde, and Rhoshel K. Lenroot | II. **Memory, Meaning, and Representation** | 2. Semantic and Associative Relations in Adolescents and Young Adults: Examining a Tenuous Dichotomy, Ken McRae, Saman Khalkhali, and Mary Hare | 3. Representation and Transfer of Abstract Mathematical Concepts in Adolescence and Young Adulthood, Jennifer A. Kaminski and Vladimir M. Sloutsky | 4. A Value of Concrete Learning Materials in Adolescence, Kristen P. Blair and Daniel L. Schwartz | 5. Higher-Order Strategic Gist-Reasoning in Adolescence, Sandra B. Chapman, Jacquelyn F. Gamino, and Raksha Anand | III. **Learning, Reasoning, and Problem Solving** | 6. Better Measurement of Higher-Cognitive Processes Through Learning Trajectories and Diagnostic Assessments in Mathematics: The Challenge in Adolescence, Jere Confrey | 7. Adolescent Reasoning in Mathematical and Non-Mathematical Domains: Exploring the Paradox, Eric Knuth, Charles Kalish, Amy Ellis, Caroline Williams, and Mathew Felton | 8. Training the Adolescent Brain: Neural Plasticity and the Acquisition of Cognitive Abilities, Sharona M. Atkins, Michael F. Bunting, Donald J. Bolger, and Michael R. Dougherty | 9. Higher Cognition Is Altered by Non-Cognitive Factors: How Affect Enhances and Disrupts Mathematics Performance in Adolescence and Young Adulthood, Mark H. Ashcraft and Nathan O. Rudig | IV. **Judgment and Decision Making** | 10. Risky Behavior in Adolescents: The Role of the Developing Brain, Adrianna Galvan | 11. Affective Motivators and Experience in Adolescents' Development of Health-Related Behavior Patterns, Sandra L. Schneider and Christine M. Caffray | 12. Judgment and Decision Making in Adolescence: Separating Intelligence From Rationality, Keith Stanovich, Richard F. West, and Maggie E. Toplak | 13. A Fuzzy-Trace Theory of Adolescent Risk Taking: Beyond Self-Control and Sensation Seeking, Christina Chick and Valerie F. Reyna | V. **Epilogue** | 14. Paradoxes in the Adolescent Brain in Cognition, Emotion, and Rationality, Valerie F. Reyna and Michael R. Dougherty

ALSO OF INTEREST

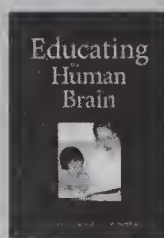


Poverty and Brain Development During Childhood

An Approach From Cognitive

Psychology and Neuroscience
Sebastián J. Lipina
and Jorge A. Colombo
2009. 172 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95
ISBN 978-1-4338-0445-8 | Item # 4318053



Educating the Human Brain

Michael I. Posner
and
Mary K. Rothbart
2007. 263 pages.
Hardcover.

List: \$39.95 | APA Member/Affiliate: \$34.95
ISBN 978-1-59147-381-7 | Item # 4318029



Child Development at the Intersection of Emotion and Cognition

Edited by
Susan D. Calkins
and Martha Ann Bell

2010. 261 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-0686-5 | Item # 4318067



AMERICAN PSYCHOLOGICAL ASSOCIATION

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

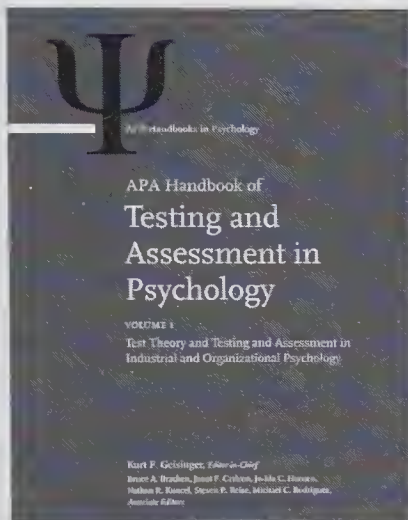
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2012

APA HANDBOOK OF TESTING AND ASSESSMENT IN PSYCHOLOGY

Volume 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology | Volume 2. Testing and Assessment in Clinical and Counseling Psychology | Volume 3. Testing and Assessment in School Psychology and Education

Editor-in-Chief Kurt F. Geisinger



This three-volume handbook is a comprehensive presentation of the theory and application of tests in psychology and education. It begins with an in-depth portrayal of psychometrics: the quantitative underpinning of testing. It then provides thorough, up-to-date and informative chapters related to five general application areas of testing: industrial/organizational psychology, clinical psychology (including health psychology), counseling psychology, school psychology and educational testing. In each of these five areas, this handbook is probably the most comprehensive review of the use of testing and assessment in the subfield.

Series: APA Handbooks in Psychology™. 2013. 2,220 pages.

Titles in the *APA Handbooks in Psychology™* Series are also available electronically to institutions via the APA PsycNET® platform.

For more information, visit <http://www.apa.org/pubs/books/institutions/handbooks.aspx> or contact the APA Licensing Department at 877-236-2941.

3-Volume Set | List: \$695.00 | APA Member/Affiliate: \$395.00 | ISBN 978-1-4338-1227-9 | Item # 4311510

CONTENTS:

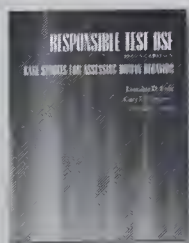
Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology | Part I: Test Theory | Part II: Types of Testing | Part III: Industrial and Organizational Psychology

Volume 2: Testing and Assessment in Clinical and Counseling Psychology | Part I: General Issues in Testing and Assessment in Professional Psychology | Part II: Clinical and Health Psychology | Part III: Counseling Psychology

Volume 3: Testing and Assessment in School Psychology and Education | Part I: School Psychology | Part II: Educational Testing and Measurement | Part III: Future Directions

To view the entire Table of Contents with contributors, please visit the book's page online at:
<http://www.apa.org/pubs/books/4311506.aspx>

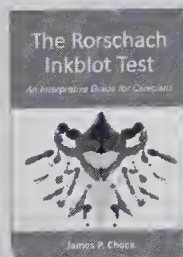
ALSO OF INTEREST



Responsible Test Use
Case Studies for Assessing Human Behavior
SECOND EDITION

Lorraine D. Eyde, Gary J. Robertson, and Samuel E. Krug
2010. 217 pages. Hardcover.

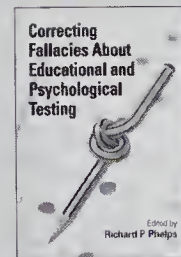
List: \$39.95 | APA Member/Affiliate: \$29.95
ISBN 978-1-4338-0556-1 | Item # 4311013



The Rorschach Inkblot Test
An Interpretive Guide for Clinicians
James P. Choca

2013. 281 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$59.95
ISBN 978-1-4338-1200-2 | Item # 4317293



Correcting Fallacies About Educational and Psychological Testing

Edited by Richard P. Phelps
2009. 287 pages. Hardcover.

List: \$39.95 | APA Member/Affiliate: \$34.95
ISBN 978-1-4338-0392-5 | Item # 4318046



AMERICAN PSYCHOLOGICAL ASSOCIATION

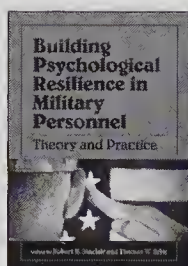
APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502
In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2249

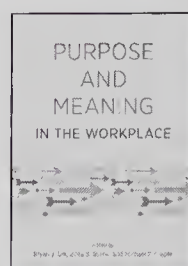
NEW RELEASES

from the American Psychological Association



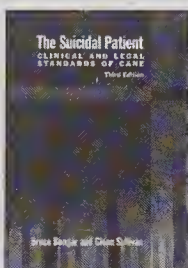
Building Psychological Resilience in Military Personnel
Theory and Practice
 Edited by Chief Robert R. Sinclair and Thomas W. Britt
 2013. 256 pages. Hardcover.

ISBN 978-1-4338-1331-3 | Item # 4317311
 List: \$69.95 | APA Member/Affiliate: \$49.95



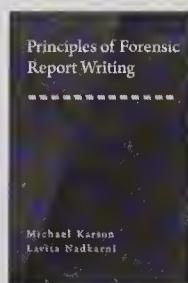
Purpose and Meaning in the Workplace
 Edited by Bryan J. Dik, Zinta S. Byrne, and Michael F. Steger
 2013. 240 pages. Hardcover.

ISBN 978-1-4338-1314-6 | Item # 4318117
 List: \$69.95 | APA Member/Affiliate: \$49.95



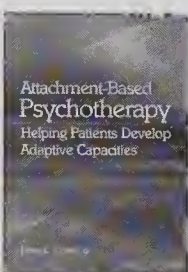
The Suicidal Patient
Clinical and Legal Standards of Care
 THIRD EDITION
 Bruce Bongar and Glenn R. Sullivan
 2013. 416 pages. Hardcover.

ISBN 978-1-4338-1325-2 | Item # 4317307
 List: \$69.95 | APA Member/Affiliate: \$49.95



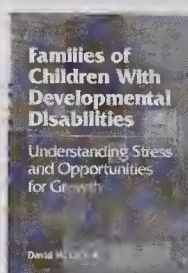
Principles of Forensic Report Writing
 Michael Karson and Lavita Nadkarni
 2013. 208 pages. Hardcover.

ISBN 978-1-4338-1306-1 | Item # 4317306
 List: \$59.95 | APA Member/Affiliate: \$49.95



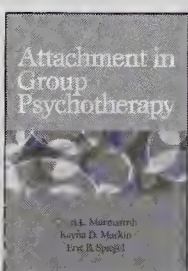
Attachment-Based Psychotherapy
Helping Patients Develop Adaptive Capacities
 Peter C. Costello
 2013. 264 pages. Hardcover.

ISBN 978-1-4338-1302-3 | Item # 4317304
 List: \$59.95 | APA Member/Affiliate: \$49.95



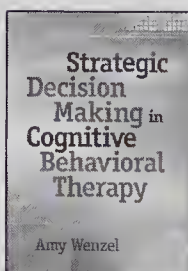
Families of Children with Developmental Disabilities
Understanding Stress and Opportunities for Growth
 David W. Carroll
 2013. 240 pages. Hardcover.

ISBN 978-1-4338-1329-0 | Item # 4316155
 List: \$59.95 | APA Member/Affiliate: \$49.95



Attachment in Group Psychotherapy
 Cheri L. Marmarosh, Rayna D. Markin, and Eric B. Spiegel
 2013. 296 pages. Hardcover.

ISBN 978-1-4338-1321-4 | Item # 4317309
 List: \$59.95 | APA Member/Affiliate: \$49.95



Strategic Decision Making in Cognitive Behavioral Therapy
 Amy Wenzel
 2013. 320 pages. Hardcover.

ISBN 978-1-4338-1319-1 | Item # 4317308
 List: \$49.95 | APA Member/Affiliate: \$44.95



AMERICAN PSYCHOLOGICAL ASSOCIATION

TO ORDER: 800-374-2721 • www.apa.org/pubs/books

AD2263